Diffusion and Supercritical Spreading Processes on Complex Networks

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)
im Fach Physik, Spezialisierung:
Theoretische Physik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
Humboldt-Universität zu Berlin

von
M.Sc. Flavio Iannelli

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:
Prof. Dr. Elmar Kulke

Gutachter:
1. Prof. Igor M. Sokolov
2. Prof. Jürgen Kurths
3. Prof. Angelo Vulpiani

Tag der mündlichen Prüfung: 20. Dezember 2018

To my family

Abstract

The large amount of datasets that became available in recent years has made it possible to empirically study humanly-driven, as well as biological complex systems to an unprecedented extent. In parallel, the prediction and control of epidemic outbreaks have become very important for public health issues. The rapid growth of transportation means, frequency of movements, web content as well as online social platforms has further increased the risk that global emergent diseases will spread worldwide or enhance fake news dissemination. The underlying networks are usually scale-free. This implies the absence of the epidemic threshold that allows pathogens and social content to easily spread in a population of individuals. On the one side, stochastic simulations of diffusive spreading, as well as more refined metapopulation models based on reaction-diffusion equations, allows us to build realist data-driven models that are a powerful tool to make detailed forecasts. On the other hand, algebraic methods give a solid foundation for drawing general conclusions and in many cases provide numerical instruments superior to direct simulations.

In this thesis, we investigate some important aspects of diffusion phenomena and spreading processes unfolding on networks. We study three different problems related to spreading processes in the supercritical regime. First, we study reaction-diffusion on ensembles of random networks characterized by the observed Lévy-flight properties of human mobility. Leveraging effective medium theory, we are able to quantitatively estimate the diameter of the infected region for a very general transportation system. The second problem is the estimation of the arrival times of global pandemics. To this end, we derive and identify suitable hidden geometries of network-driven spreading processes, leveraging on random-walk theory. Through the definition of network *effective distances*, the problem of complex spatiotemporal patterns is reduced to simple, homogeneous wave propagation patterns. Third, by embedding nodes in the hidden space defined by network effective distances, we introduce a novel network centrality, called *ViralRank*, which quantifies how close a node is, on average, to the other nodes. As a case study, we first characterize the political leanings and, using known heuristic centralities, rumor spreading dynamics on two networks built on datasets extracted from Twitter on the specific topic of the 2016 constitutional referendum in Italy. Then, we investigate the role of centrality measures in identifying influential spreaders by comparing the relative performance with ViralRank in several empirical datasets of social, biological and infrastructure complex systems. We find that ViralRank can correctly identify influential nodes in the supercritical regime for both contact networks and metapopulations, as it systematically outperforms state-of-the-art centrality measures. Our results bring us closer to the optimal solution to the problem of the influential spreaders identification. These three studies constitute a unified framework to characterize diffusion and spreading processes unfolding on complex networks in very general settings, and provide new approaches to challenging theoretical problems that can be used to benchmark future models.

Keywords: Complex Networks, Epidemics, Influencers, Statistical Physics

Zusammenfassung

Die große Menge an Datensätzen, die in den letzten Jahren verfügbar wurden, hat es ermöglicht, sowohl menschlich-getriebene als auch biologische komplexe Systeme in einem beispiellosen Ausmaß empirisch zu untersuchen. Parallel dazu ist die Vorhersage und Kontrolle epidemischer Ausbrüche für Fragen der öffentlichen Gesundheit sehr wichtig geworden. Die Entwicklung schnellerer Transportmittel und deren häufigere Benutzung sowie ein rasant wachsendes Internet und soziale Online-Medien haben das Risiko, dass sich Krankheiten sowie Falschmeldungen weltweit verbreiten, weiter erhöht. Die zugrunde liegenden Netzwerke sind in der Regel scale free. Dies impliziert das Fehlen der epidemischen Schwelle, was Pathogenen und sozialen Inhalten erlaubt, sich in einer Population von Individuen leicht zu verbreiten. Auf der einen Seite erlauben uns stochastische Simulationen der diffusiven Ausbreitung sowie Metapopulationsmodelle auf der Grundlage von Reaktions-Diffusions-Gleichungen, realistische Datenmodelle zu erstellen, die ein leistungsfähiges Werkzeug für detaillierte Vorhersagen sind. Auf der anderen Seite bilden algebraische Methoden eine solide Grundlage für allgemeine Schlussfolgerungen und liefern in vielen Fällen numerische Instrumente, die direkten Simulationen überlegen sind.

In dieser Arbeit untersuchen wir einige wichtige Aspekte von Diffusionsphänomenen und Ausbreitungsprozessen auf Netzwerken. Wir untersuchen drei verschiedene Probleme im Zusammenhang mit Ausbreitungsprozessen im überkritischen Regime. Zunächst untersuchen wir die Reaktionsdiffusion auf Ensembles zufälliger Netzwerke, die durch die beobachteten Lévy-Flugeigenschaften der menschlichen Mobilität charakterisiert sind. Mit Hilfe der Effektive-Medium-Theorie können wir den Durchmesser der infizierten Region für ein sehr allgemeines Transportsystem quantitativ abschätzen. Das zweite Problem ist die Schätzung der Ankunftszeiten globaler Pandemien. Zu diesem Zweck leiten wir geeignete verborgene Geometrien netzgetriebener Streuprozesse, unter Nutzung der Random-Walk-Theorie, her und identifizieren diese. Durch die Definition von *effective distances* wird das Problem komplexer raumzeitlicher Muster auf einfache, homogene Wellenausbreitungsmuster reduziert. Drittens führen wir durch die Einbettung von Knoten in den verborgenen Raum, der durch effective distances im Netzwerk definiert ist, eine neuartige Netzwerkzentralität ein, die *ViralRank* genannt wird und quantifiziert, wie nahe ein Knoten, im Durchschnitt, den anderen Knoten im Netzwerk ist. Als Fallstudie charakterisieren wir zunächst die politischen Neigungen und, unter Verwendung bekannter heuristischer Zentralitäten, die Dynamik von Gerüchten in zwei Netzwerken, die auf Daten basieren, welche aus Twitter zum spezifischen Thema des Verfassungsreferendums 2016 in Italien extrahiert wurden. Anschließend untersuchen wir die Rolle von Zentralitätsmaßen bei der Identifizierung einflussreicher Streuer durch den Vergleich der relativen Leistung mit ViralRank in mehreren empirischen Datensätzen sozialer, biologischer und infrastruktureller komplexer Systeme. Wir stellen fest, dass ViralRank sowohl für Kontaktnetzwerke als auch für Metapopulationen einflussreiche Knoten im überkritischen Regime korrekt identifizieren kann, da systematisch herkömmliche Zentralitätsmaße übertroffen werden. Unsere Ergebnisse bringen uns der optimalen Lösung, für das Problem der Identifizierung einflussreicher Streuer, näher. Diese drei

Studien bilden einen einheitlichen Rahmen zur Charakterisierung von Diffusions- und Ausbreitungsprozessen, die sich auf komplexen Netzwerken allgemein abzeichnen, und bieten neue Ansätze für herausfordernde theoretische Probleme, die für die Bewertung künftiger Modelle verwendet werden können.

Schlagwörter: Komplexe Netzwerke, Epidemiologie, Einflussreiche Streuer, Statistische Physik

Acknowledgements

First of all, I want to thank Igor Sokolov for giving me the opportunity to work with him on such vibrant and exciting topics and also for being the silent but always present guide that I needed to develop the *know-how* at the early stage of my research career. Thanks to his experience and wiseness I was able to learn independently the skills necessary to make all the work that culminated in this thesis. Working with him has been a very important experience for me thanks to his scientific expertise and guidance as well as the ability to create a stimulating and comfortable working environment. My hope is therefore to be able to continue to work with him in the future.

I also sincerely thank my previous mentors Giorgio Parisi and Massimo Testa, who taught me theoretical physics when I was a student at "*La Sapienza*" in Rome and Angelo Vulpiani for giving me valuable advice that directed me toward statistical physics later on. My enormous gratitude goes to all my collaborators who shared with me the constant curiosity and enjoyed the ride to the unknown. For this reason I wish to thank Dirk Brockmann, Philipp Hövel, Andreas Koher, Manuel Sebastian Mariani, Felix Thiel and the "Italian gang": Jacopo Bindi, Davide Colombi, Nicola Politi, Michele Sugarelli, Raffaele Tavarone and Enrico Ubaldi.

During my PhD I had the opportunity to travel to various conferences, workshops and schools and to interact and exchange ideas with many experts and leaders in the field of network science. In particular, I thank Albert-László Barabási, Vittoria Colizza, Diego Garlaschelli, Shlomo Havlin, Samuel Johnson, Romualdo Pastor-Satorras, Sidney Redner and Vinko Zlatić for many interesting discussions.

I will never forget the first school at the very beginning of my PhD and all the great adventures riding volcanoes with Manuel in Lipari, especially thanks to the incessant laughs with Danilo "il birrista" Leuzzi and Matteo "l'avvocato" Morini. I also thank Lyuba and Dima for the fantastic hospitality in Moscow and for showing the great CCEGN group the best way to drink Russian vodka, and also for setting up a conference with such an amazing group of people.

All these years spent at HU Physik have been great fun particularly thanks to my office

mates Felix, Stephan and later Anna – who finally became an expert in the problem of the square in quantum mechanics – and also to all other 3rd floor members: Bernard, Chris, Justus, Martin, Mohsen, Patrick, Paul, Stan and later Fabian. For the same reason I am proud to be part of the last IRTG group at HU with Jörg "Don George" Nötel and Malte "Don Maltche" Kaehne, who definitely made the three years plus working here and the Brazilian workshops as fun as it can possibly get. I especially thank David who was more than just the secretary of the group and who was always there for anything and for all of us. All afternoons spent at TU Physik have been also a lot of fun, thanks to Andreas, Jason "Giasone il greco" and later Philipp.

The work culminated in this thesis has been carried out between Berlin and the state of São Paulo, Brazil. I acknowledge many fruitful discussions there with Tiago Pereira, Leonardo Santos, and Didier Augusto Vega-Oliveros. I thank all the fantastic people that I met there: Anderson, Pedro and Tiago for the great hospitality, and Sabrina and Ian for the unforgettable time spent together in Sao José dos Campos, in Ilhabela and in the Brazilian countryside, and the fabulous INPE basketball team. I also would like to thank the "almost Brazilian" Franziska that shared with me the first part of the adventure in Brazil listening to Bombino on the bus towards paradise and Julio for being the best guide for me and Stephan through the night jungle of São Paulo. All those days away from home as well as the routine mornings and evenings in the S-Bahn and through the Adlershof fields and construction sites could not have been the same without Boards of Canada.

Finally, I want to thank my parents, Giulio, who is becoming a much wiser physicist than I could possibly hope to become and Anna and Flora for having the necessary patience to deal with me every single day and for sharing with me all this.

Contents

List of Symbols and Abbreviations

G	Graph: A tuple $G = (\mathcal{V}, \mathcal{L})$ of a set of vertices (nodes) $\mathcal{V}$ and a set of links (edges) $\mathcal{L}$
$\mathcal{V}$	Set of vertices
$\mathcal{L}$	Set of links
N	Number of nodes of a network, given by the cardinality of the set of vertices $\mathcal{V}$
E	Number of edges of a network, given by the cardinality of the set of links $\mathcal{L}$
D	Diameter of a network
$\langle D \rangle$	Average shortest-path length of a network
$\langle C \rangle$	Average (global) clustering coefficient of a network
$\mathbf{A}$	Adjacency matrix (unweighted)
$\mathbf{W}$	Weighted adjacency matrix
$\mathbf{I}$	Identity matrix
$\mathbf{E}$	Matrix of ones
$\mathbf{L}$	Laplacian matrix
$\mathbf{P}$	Transition probability matrix
$\mathbf{Q}$	Transition rate matrix
$\mathbf{M}$	Mean-first passage time matrix
$\mathbf{H}$	Hitting-time probability matrix
k_i	Degree of a node
$k_i^{out},\ k_i^{in}$	Out-degree and in-degree of a node
s_i	Strength of a node
q_i	Exit rate of a node
e_i	Vector of ones
π_i	Stationary density vector

$\mathcal{D}$	Diffusion coefficient
Γ	Path in a network
Ξ	Walk in a network
α	Diffusion rate
β	Transmission rate
μ	Recovery rate
$\mathcal{R}_0$	Basic reproductive number

RL	Regular lattice
ER	Erdös-Rényi
WS	Watts-Strogatz
BA	Barabási-Albert
MF	Mean field
$MFPT$	Mean-first passage time
EMT	Effective medium theory
ED	Effective distance
$RWED$	Random-walk effective distance
RN	Retweet network
MN	Mention network
FJ	Friedkin-Johnsen
$WCGC$	Weakly connected giant component
$SCGC$	Strongly connected giant component
SI	Susceptible-infected
SIS	Susceptible-infected-susceptible
SIR	Susceptible-infected-removed
ISS	Ignorant-spreader-stifler

List of Figures

List of Tables

1

Introduction

> "*The central task of theoretical physicists in our time is no longer to write down the ultimate equations but rather to catalog and understand emergent behavior in its many guises, including potentially life itself*"
>
> –David Pines, *The Theory of Everything*

Network science is a relatively new field of research that has become synonym with the study of *complex systems*. Indeed the increasing level of attention that the study of networks has been receiving is due to their broad applicability in describing a wide range of different phenomena. Prominent examples are the mapping of the World Wide Web and structure of the physical Internet [217], economic and financial systems [27], social and language dynamics [59], transport and human mobility [164], altered state of consciousness in the human brain [224] and even the cosmic web of galaxies [80], see Figure 1.1.

Although no definition of complex system is universally accepted and varies across disciplines, all complex systems are generally characterized by properties emerging from the interactions between a large number of components. A complex system differs from a simple system in that microscopic and macroscopic scales cannot be treated separately, in contrast to the Newtonian paradigm whereby the world is reducible to a few fundamental elements leading to predictable behavior. Complexity emerges not just as an excellent

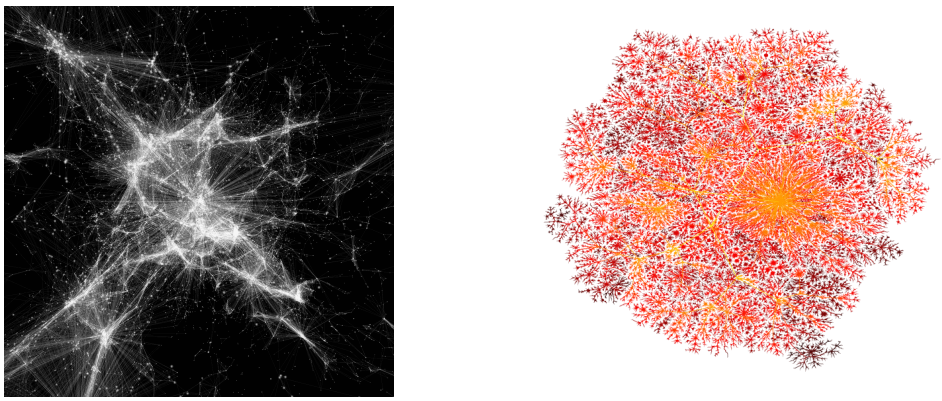


Figure 1.1: Left panel: Network visualization of the cosmic web produced by a varying length model, where the length of each connection is proportional to the size of the connected galaxies [80]. (Credit: Courtesy of Kim Albrecht). Right panel: Topology of the Internet of autonomous systems at the end of the 20th century, produced by the Cooperative Association for Internet Data Analysis (CAIDA) within the Internet Mapping Project (Credit: Courtesy of William Cheswick).

way to put certain intriguing concepts, but as a phenomenon that is deeply rooted in the laws of Nature [203].

The fundamental approach to understand how to connect the microscopic to the macroscopic was formalized in the nineteen century through statistical mechanics and condensed matter physics. The methods of statistical physics can then be applied to understand and interpret the qualitative and quantitative behavior of complex systems such as amorphous materials, strongly disordered systems (glasses), collective animal behavior, socioeconomic and biological systems, chaotic oscillations and the human brain [86, 267, 203, 251].

In his famous article [6], Anderson critically addressed the reductionist hypothesis of science. The task of reconstructing how the universe works by adding together very simple physical laws breaks down when confronted with large aggregates, where entirely new properties appear and different scales of complexity emerge: *More is different.* Complex structures arise in Nature even in simple situations and the observed complexity is very often contrasted with the astonishing simplicity of the basic laws of physics. One way to define complexity is to identify it by structure with variations [128], as each complex system is different. Some have also stressed the importance for a complex system of being out-of-equilibrium and to self-organize [204]. However, the complexity paradigm is probably best understood in terms of the conceptual problems of critical phenomena and in particular second order phase transitions. At the critical point, fluctuations are important on *all* length scales connecting the microscopic with the macroscopic and assume a self-similar structure. The same ideas of self-similarity are found in the study of

fractal growth phenomena [19, 52] and the tendency of complex systems to self-organize to a critical state with self-similar properties has been the object of many studies, e.g. in models of diffusion-limited-aggregation (DLA) [277]. In the context of networks it turns out that, growth together with some simple requirement can give rise to very complex structures [18], with the same self-similar properties of systems at criticality [250, 248]. As for the critical point of phase transitions, a trademark of complexity is the power-law behavior of some quantity that characterizes the system.

Power-law distributions are indeed abundant in Nature [199]. Meteorite sizes, city sizes, income and the number of species per genus follow power-law distributions [244]. In the context of animal movements such power laws have been observed in foraging movements of many species [269, 26, 190, 268, 230, 10]. Power laws are also found when analyzing the density distributions and velocity fluctuations of starling flocks [60] and even in the distribution of connections in the causal network representing the large-scale structure of spacetime in our accelerating universe [168].

One common misconception is to identify complex with complicated [23]. However, the distinction between the two is a crucial one and the behavior of a complex system is generally different and richer than an analogous complicated system. The characteristics of emergent behavior, self-organization and self-similarity in their structure generally characterize a complex system as the spontaneous outcome of the interactions among the many constituent units. Contrary, even very complicated systems made of a large number of constituents that are engineered and put in place according to a definite blueprint, lack all these distinctive features [23]. An additional common feature of many complex systems is resistance to random removals of their components, while this would lead rapidly to the total failure of merely complicated systems.

In the case of networks, the variation at all scales is statistically encoded in the heavy-tail distributions characterizing the structural properties. Then, the larger the size of a system, the more significant its heterogeneity with fluctuations extending over all the orders of magnitude allowed. With virtually infinite fluctuations, it is then impossible to define a typical scale in which an average description would be reliable. The evidence that a complex topology is the ubiquitous outcome of the evolution of networks can be hardly considered as incidental [23], but rather *universal*. These ideas are also at the core of the complexity pyramid in living organisms [208], characterized by the gradual transition from the particular (at the bottom level) to the universal (at the apex). Indeed, although the individual components are unique to a given living organism, the topological properties of different cellular networks share surprising similarities with those of natural and social networks.

We can argue that behind each complex system there is a network [17]. For this reason, we will never understand complex systems unless we map out and understand the networks behind them. This is particularly relevant since in Nature everything is simple, except, of course, Nature itself [128]. Thus, the missed identification of the proper network representation impairs our ability to use network theory successfully. Network

science is powerful because structure determines function and the way we assign the connections determines how the content of a system manifests itself.

In this work we study dynamical process on *complex networks*, and in particular we investigate *diffusion* and *spreading* phenomena. Epidemic spreading in humans and animals as well as social contagion in virtual platforms are ubiquitous phenomena in our society. The epidemic modeling metaphor has been introduced to describe a wide array of different phenomena [214]. Among others, the spread of information and cultural norms, how blackouts spread nationwide or how efficiently memes can spread on social networks can all be conceptually modeled as a contagion process, whose mathematical description is built on models similar to classic epidemic models. Although the detailed mechanisms of each phenomenon can be very different, on a coarse-grained level their mathematical description is often framed by the constitutive equations of the general theory of reaction-diffusion processes [262].

From a physicist perspective, spreading processes belong to the class of *non-equilibrium critical phenomena*, characterized by a crossover between an active and an absorbing phase. Contrary to equilibrium phase transitions, the stationary state of the system is not an equilibrium state and is characterized by lack of reversibility in its dynamics. This manifests itself as the breaking of detailed balance and prevents us from using the theoretical framework of equilibrium statistical mechanics, where the statistical weight depends only on the specific static configuration and not on the whole history.

The standard way to study biological contagion is through compartmental models where one divides the population into compartments describing the state of each component. This simplified approach that assumes homogeneously mixed populations [139], is generalized to the much more realistic scenario in which the detailed structure of the interaction network is considered and agents can only spread through the given connections. In social contagion ideas spread along social networks in a manner similar to biological contagion and can become *viral*. Further generalizations in the social context are done considering that unlike biological contagion, ideas spread in a manner that involves social reinforcement, leading to so-called complex contagions [273, 61]. For both biological and social contagions, the active and absorbing phases are separated by an absorbing critical point that defines the *epidemic threshold* of the model. The concept of epidemic threshold is very general and applies to very different epidemic models [139]. Processes that are in the active phase (above the threshold) are called *supercritical*, and *subcritical* otherwise.

Network-mediated spreading processes are ubiquitous. Online users transmit news and information to their contacts in online social platforms [14, 219, 284], individuals form their opinion and make decisions influenced by their contacts in social networks [88, 114, 112] and infected individuals can transmit infectious diseases to their sexual partners [97]. Networks constitute the substrate for the spreading of agents as diverse as computer viruses [156, 157], deadly pathogens [7] and rumors [185, 84]. Crucially, real-world networks of relevance for epidemic spreading are different from regular lattices.

Complex networks are hierarchically organized with a few nodes that may act as hubs and where the vast majority of nodes have few interactions. Fortunately, today we are able to measure the space of spreading processes through modeling and a great amount of available data, both for the structures (the interaction networks) on which such processes evolve, as well as for the specific epidemics on the biological side. Obviously, even a single infection event is an infinitely complex process which is impossible to model. To obviate the problem, one usually adopts a *coarse-grained* description of the dynamical process. The computer as the "fast abacus" [181] then becomes the laboratory where models can be run to create *in silico* experiments that would be infeasible in real systems. Numerical simulations become the creator of the phenomena that we want to study.

At the end of the last century and especially in very recent years, physicists became interested in socioeconomic problems also driven by the large availability of data. Besides, the emergence of this "physics of data" [53] led to the development of *econophysics* [189] and *sociophysics* [222]. Very complex phenomena such as economic growth, technological development and opinion formation can be understood applying novel approaches empirically grounded on *Big Data* analysis, e.g. to unravel the pattern of economic development and technological innovation [140]. Prominent examples of the success of this approach are the theory of *economic complexity* [141] and the related novel algorithms for the forecasting of Gross Domestic Product (GDP) growth [254, 83], which explains how the product space of nations shapes the macroeconomic growth. The recent trends in sociophysics are related to so-called computational social science, that relies on a data-driven approach to studying social phenomena [245]. These data contain information on what people do when using different services on mobile devises such as search engines, online banking and social networks. The wide availability of user-provided content in online social media facilitates the aggregation of people around common interests, narratives and political leanings. Besides, this also allows for the rapid dissemination of unsubstantiated rumors and conspiracy theories [84, 90].

The common theme of social dynamics is the understanding of the transition from an initially disordered state, to a configuration that displays order. Modeling the transition between order and disorder is common in statistical physics and is formalized by the Ising paradigm [59]. Opinion dynamics in humans or even cooperative transport in groups of ants [102] can easily be modeled this way. Individuals are assigned an opinion (spin) that can switch between positive and negative value by interacting with their neighborhood. A recurrent criticism on this approach is that the entities that represent individuals, such as the nodes in a network, can barely be captured by up and down spins. Successful sociophysics models bridge the micro and the macro following principles of data-driven modeling [245] and are validated by a quantitative comparison of the simulated dynamics with real observations.

To predict the behavior of a large number of interconnected techno-social systems, it becomes a necessity to start with the mathematical description of patterns found in real-world data. The modern approach to epidemic modeling that is evolving by

the day, both for technology advancement and data acquisition, is becoming close to weather forecast for diseases. The basic difference with weather forecasting, where we know the physical laws governing fluids and gasses, is that for techno-social systems the modeling is inherently made harder by the very limited knowledge of society and human behavior [264]. Analogously to what happened in physics with the shift from atomic and molecular physics to condensed matter, today the large amount of available data allows us to study quantitatively the behavior of large aggregates of "social atoms".

The goal of epidemiology is to understand the patterns of disease and health dynamics in populations as well as the causes of these patterns, and to use this understanding to mitigate and prevent large scale outbreaks. Digital epidemiology [238] has emerged in the past few years as a new field driven by the increasing data availability and computing power, as well as by breakthroughs in data analytics methods. The abundance of data in recent years combined with the network approach is the key ingredient of modern epidemic modeling. This is changing dramatically our understanding of a wide range of phenomena emerging from the interplay between epidemic processes and networks [214].

Understanding the spread of emergent infectious diseases in the geographic space is particularly important in an increasingly interconnected world [133, 239]. In ancient times, the spreading of epidemics such as the Black Death [197], could be understood in terms of a spatial diffusion phenomenon [17]. In those cases the disease is spread by the individuals that can only travel with low velocities bounded by the local connectivity of the geographical space. This gives rise to a *wave-front* of infected individuals, which travels at a finite speed. Contrary, modern transportation networks are characterized by large fluctuations in the connectivity among densely populated areas and the correspondent urbanization. Furthermore, the complexity of human mobility at all scales [48, 129], being that urban and inter-urban or world-wide, is reflected in the possibility for the infection to cross arbitrary distances in close to no time. As a consequence, the epidemic prevalence quantified by the number of infected sites grows exponentially fast, as opposed to linearly. Similar phenomena are also discussed in the biological context [132]. Emergent epidemic threats such as H1N1 [279], SARS [73] or EBOV [226], and more recently ZIKV [151], make the prediction and control of global epidemic outbreaks a central task for public health issues [225, 145]. The large amount of traffic data both at the local and global scale provides a new opportunity to understand such processes.

On the one hand numerical simulations of infection spreading offer a practical tool for estimating key epidemic quantities such as the *infection arrival time* [49]. Mathematical models of spreading can be studied based on two different frameworks. At the local scale, *contact-network* models of spreading assume that individuals directly infect the individuals they are in contact with. The topology of the underlying network of contacts plays a critical role in determining the size of the infected population [215]. Instead, at the global scale *reaction-diffusion* models assume that individuals can infect the individuals that belong to the same population (reaction process), and infected individuals can move across populations (diffusion process). This *metapopulation* approach is increas-

ingly used to forecast the properties of epidemic outbreaks [175, 74, 15, 11, 261], and to design and understand the systemic impact of disease containment strategies [258]. On the other hand, algebraic methods give a solid foundation for drawing general conclusions and in many cases provide numerical instruments superior to direct simulations.

Numerical models allow us to investigate the fundamental problem of identifying those nodes which, once they initiate a spreading process, maximize the size of the infected population [159, 85, 29, 228]. Identifying such nodes, commonly referred to as *influential spreaders*, is vital for organizations to design effective marketing campaigns in order to maximize their chances of success [92, 155, 176], for policy-makers to design effective immunization strategies against infectious diseases [71], for social media companies to maximize the outreach of a given piece of information, such as a news or a meme [41]. The identification of influential spreaders is benchmarked by running multiple realizations of spreading models on real networks, with different "seed" nodes as initiators of the process. The typical size of the outbreak generated by each seed node quantifies its ground-truth *spreading ability*. One can thus compare different strategies for assigning a score to nodes based on centrality measures, with respect to their ability to identify the nodes with the largest ground-truth spreading ability [182].

In this thesis we present a throughout investigation of diffusion and spreading processes on complex networks. Three important aspects of epidemic spreading for both biological and social contagions are analyzed in detail. First, we consider epidemic spreading on very general transportation networks at the global scale, by constructing artificial random networks with spatial embeddings. The mathematical form of the intensity of each connection is chosen in order to model the characteristic scale-free motion of observed human mobility [48, 129]. By leveraging effective medium theory, a framework to evaluate disorder averages of random networks, we extract relevant epidemiological quantities in spatially embedded metapopulations with long-range connections. Second, we derive an analytical network-based measure built on random-walk hitting times, called *effective distance*. Three different approaches to define effective distances are discussed in detail: (i) the dominant-path, (ii) the multiple-path and (iii) the random-walk approach. Using a microscopic description of the spreading process, we are able to bridge concepts of epidemic spreading in structured populations with random walks on networks, by leveraging the mathematical formalism of extreme event statistics [131]. The random-walk effective distance that we define, which has a clear interpretation and is computationally feasible for large networks, is able to reduce complex spatiotemporal patterns to simple, homogeneous wave propagation patterns. Contrary to previous attempts, based on the dominant-path approach, that can significantly overestimate the numerical infection arrival time, we are able to quantify with very high precision the spreading patterns in the hidden geometry induced by effective distances. To validate our analytics, we use a comprehensive dataset of global mobility, obtained from the Official Airline Guide. The third and last study is devoted to the problem of identification of influential spreaders. As a case study, we first analyze opinion dynamics and social contagion on the online

social platform Twitter. For this purpose we download user posts (tweets) on the specific topic of the 2016 constitutional referendum in Italy. By leveraging machine learning techniques, we develop an analytical framework to assign dynamical opinions to users based on the content of their activity in the three months prior to the referendum vote. From this procedure we construct two temporal networks, one of users mentions and one of content retweet. We find that, the global opinion averaged over all users is in very good agreement with official pool statistics and that the final result of the referendum is well reproduced by the mathematical framework that we chose to assign users' opinions during the political debate. Then, we simulate numerically a rumor spreading in the two networks that we constructed and rank users by their spreading ability. By comparing heuristic nodes' centrality measures, we find that the number of connections (mentions or retweets) of each user can provide an extremely accurate description of the rumor spreading ability in the political discussion on Twitter. Next, we introduce a new metric called *ViralRank*, by embedding the nodes of a network in the hidden space defined by its random-walk effective distances. By comparing the correlation between scores of the nodes assigned by the ground-truth spreading ability and state-of-the-art centralities, we find that ViralRank systematically outperforms known methods in the supercritical regime, when the spreading process reaches a substantial portion of the network. In addition we find that our measure can be expressed in terms of a known opinion formation model, devised for modeling the reach of consensus in real social experiments. Through the definition of ViralRank in an analogy with statistical mechanics, we also allow for a new and insightful interpretation of the well known Google web-ranking algorithm PageRank.

The next Chapters are organized as follows. In Chapter 2, we lay down the mathematical formalism that stands as the reference for all subsequent Chapters. In Chapter 3, we study reaction-diffusion processes in ensembles of random networks and provide an analytical expression for the epidemic growth rate. In Chapter 4, we derive network effective distances, and use them quantify the infection arrival times of reaction-diffusion processes on the global mobility network of air-traffic. In Chapter 5, we study the dynamics of opinion shifts and political leanings on the social network Twitter and identify the most influential spreaders of rumors using heuristic centralities. In Chapter 6, we introduce ViralRank, a novel network measure for nodes that outperforms state-of-the-art centralities in identifying the influential spreaders. Finally, in Chapter 7 we give a summary and outline future perspectives.

2

Dynamical Processes on Complex Neworks

"*The whole is more than the sum of its parts*"

–Aristotele, *Metaphysica 1045a*

Contents

Dynamical processes unfolding on networked structures are at the very core of this work. This Chapter is intended as the general reference and theoretical

framework for all subsequent Chapters. First, we build the basic language of graph theory necessary to characterize the properties of random and empirical networks and the mathematical formalism of dynamical processes on such networks. After defining the fundamental network models in Section 2.1, we define simple diffusion and random walks on graphs in Section 2.2. Then we analyze compartmental models of epidemic spreading in the broader context of non-equilibrium critical phenomena in Section 2.3 and conclude with a description of metapopulation models directly built on reaction-diffusion equations.

2.1. From graphs to complex networks

In this Section we review the basic notions of graph theory required to understand the next Chapters and provide a reference point for the reader, if necessary. From a physicist perspective, a graph can be thought of as a direct generalization of the regular lattice used to describe the structure and properties of matter [9]. Graph theory in its modern formulation traces back to Leonhard Euler who introduced for the first time the notion of graphs. Euler was interested in finding out if from the center of the city of Königsberg in Russia it would possible to walk crossing all seven bridges of the city only once? The fundamentally novel step forward made by Euler was to reduce the problem to a map where the geographical distances do not matter any more. Different parts of the city are described by points, called nodes, and if they are linked (by a bridge) there is a line, called an edge, between them. Through this formalism, the original problem now translates into the request of finding a *path* that passes through all the edges exactly once.

2.1.1. Graph theory in a nutshell

A graph consists of a pair of sets $G(\mathcal{V}, \mathcal{L})$, the vertices (nodes) $\mathcal{V} = \{i\}$ and the links (edges) $\mathcal{L} = \{(i, j)\}$, where (i, j) is the link from i to j. The number of nodes $N = |\mathcal{V}|$ is the *order* of the graph and the number of edges $E = |\mathcal{L}|$ its *size*. The structure of $G(\mathcal{V}, \mathcal{L})$ is represented by the $N \times N$ adjacency matrix $\mathbf{A}$, defined as

$$A_{ij} = \begin{cases} 1 & \text{if } (i, j) \in \mathcal{L} \\ 0 & \text{if } (i, j) \notin \mathcal{L} \end{cases} \tag{2.1}$$

For *undirected graphs* the associated adjacency matrix is symmetric. A simple graph of order N is an undirected graph with no self-edges and no weights associated to the edges. In this case the maximum number of edges $E_{max} = \binom{N}{2}$ is given by half[1] the

[1]For undirected graphs we must neglect the reversed connections to avoid double counting the edges.

product between any starting point N, with all possible $N-1$ remaining destinations, while $E_{min} = N-1$ is the number of edges present in a close cycle. If one accounts for self edges these must be added to get $E_{max} = N(N-1)/2 + N = N(N+1)/2$. For a directed graph (digraph) we have $E_{max} = N(N-1)$ and $E_{max} = N^2$ without and with self edges respectively.

If instead of giving only ones and zeros as entries of the adjacency matrix we consider any non-negative real number[2], we obtain a weighted graph described by the weighted adjacency matrix $W_{ij} \geq 0$. This generalization is often useful to characterize systems where the flow of information needs to be quantitatively taken into account. In this case the topology is still characterized by the unweighted adjacency matrix A_{ij}, while the weight W_{ij} represents a physical property of the edges such as the traffic volume, capacity or intensity of the interaction between pairs. Quite generally every real system that admits an abstract mathematical description as an unweighted or weighted graph is a *network* if the elements of the system (nodes) interact with each other via the edges.

A key issue regarding the structure of networks is the reachability of nodes, i.e. the possibility of moving from one node to another along the existing edges of the graph. A simple graph is connected if every node is reachable from any other node. To formalize this idea we introduce the concept of path Γ of length $n(\Gamma) = |\Gamma|$. On the graph $G(\mathcal{V}, \mathcal{L})$ a path Γ_{i_0,i_n} is defined as an ordered collection of $n+1$ nodes $\{i_0, i_1, \ldots, i_n\} \in \mathcal{V}$ and n edges $\{(i_0, i_1), (i_1, i_2), \ldots, (i_{n-1}, i_n)\} \in \mathcal{L}$. An undirected graph where any two nodes are connected by exactly only one path is called a *tree*. Importantly, no node can appear twice in the sequence of edges constituting a path. When this happens one extends the previous definition of path Γ to a *walk* Ξ. In this case a very elegant relation with the adjacency matrix exists, namely that the matrix element $(\mathbf{A}^n)_{ij}$ equals the number of walks of length $n(\Xi) = |\Xi|$ existing between node i and node j. We will always assume, unless otherwise stated, that the graph is *connected*, i.e. that there exists a path connecting any two nodes. A graph is *disconnected* if and only if there is a permutation of the indices such that the adjacency matrix can be rearranged into a block diagonal form, such that all nonzero elements in the matrix are contained in square blocks along the matrix diagonal and all other elements are zero. Each of these square blocks corresponds to a *component*. A component of a graph is defined as a connected subgraph, and the *connected giant component* is that whose size scales as the graph order, diverging in in the *thermodynamic limit* $N \to \infty$. For directed networks the situation is more complicated because some nodes are reachable in one direction (given by the edges arrows) but the reverse travel may be prohibited. In general the component structure of a directed network is divided in *weakly connected giant component* (WCGC) and *strongly connected giant component* (SCGC), considering the edges as undirected and directed respectively.

[2]In our discussion, although in principle allowed, we will always discard the possibility for negative weights.

Given a path Γ_{ij} connecting node i with node j, the geodesic (or chemical) distance D_{ij} is the minimum number of edges needed to get from i to j

$$D_{ij} = \min_{\{\Gamma_{ij}\}} \sum_{(k,l)\in\Gamma_{ij}} A_{kl}, \tag{2.2}$$

where the minimum is taken over all possible paths between the node pair. For weighted graphs we need to replace the adjacency matrix A_{kl} with the weighted adjacency matrix W_{kl} or with the reciprocal $1/W_{kl}$, in cases where higher weights correspond to lower distance. The maximum over all node pairs of (2.2) gives the diameter $D = \max_{(i,j)} D_{ij}$, i.e. the largest distance in the graph. Alternatively, the average shortest path length $\langle D\rangle = \sum_{(i,j)} D_{ij}/E_{max}$ is commonly used as a definition for the linear size of the graph.

Besides the different measures on node pairs, individual nodes can be characterized by the structure of their local neighborhood. The property that quantifies the tendency observed in real networks to create local clusters, or simply to form triangles, is the the *clustering coefficient* c_i defined as the ratio $c_i = 2e_i/k_i(k_i-1)$ between the number of edges $e_i = \sum_{j,k} A_{ij}A_{ik}A_{jk}$ of the neighbors of i and the maximum number of those edges. A graph can then by classified in terms of the *global clustering coefficient* $\langle C\rangle = \sum_i c_i/N$.

As a generalization of the adjacency matrix (2.1) we can consider multiple edges, yielding a multigraph (or multiplex network [160]), where each entry of the matrix gives the number of edges (of the same type) between that node pair. A similar idea leads to the concept of *temporal network* [143], where we have an ordered set of adjacency matrices $\{\mathbf{A}^{(t)}\}$, one for each time snapshot $t = 1,\ldots,T$. A temporal network differs substantially from a multigraph in that the order in which edges appear in the observation time matters and is responsible for the very different dynamics that can unfold on such structures. Although practically all real-world networks are described by a temporal network we can often disregard the temporal rearrangement of the edges, depending on the particular dynamical process of interest. Quantitatively, we might define the ratio τ between the network evolution time scale and the dynamics time scale and be in one of three scenarios, in analogy with disordered systems nomenclature [194]: (i) the annealed regime when $\tau \ll 1$, (ii) the intermediate regime $\tau \approx 1$ and (iii) the quenched regime $\tau \gg 1$. Only in the quenched regime we can neglect the temporal effects and study the dynamics using the standard description given by the static adjacency matrix (2.1). In the annealed regime the network evolves on a much faster time scale and the dynamics effectively occurs on the time-averaged (aggregated) network

$$\overline{\mathbf{A}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{A}^{(t)}. \tag{2.3}$$

The interesting case is the intermediate regime where, although several attempts have been put forward, a general analytical framework is still currently lacking.

2.1.2. Centrality measures

To characterize the importance of individual nodes is one of the most challenging problems in graph theory. We can define the centrality of a node using different criteria [201]. The most basic centrality is the *degree*[3]. For undirected graphs the degree k_i of node i is the number of its edges

$$k_i = \sum_j A_{ij}, \tag{2.4}$$

from which the total number of edges is computed as $E = \sum_i k_i/2 = \sum_{i<j} A_{ij}$. The discrete degree distribution[4] is the normalized histogram $P(k) = N_k/N$, where N_k is the number of nodes with degree equal to k. The average degree $\langle k \rangle = \sum_k kP(k)$ of undirected graphs is $\langle k \rangle = \sum_i k_i/N = 2E/N$. Then the *graph density*, i.e. the fraction of exiting edges with respect to the full graph, is $\rho = E/N^2 = \langle k \rangle /(2N)$. For a digraph the degree splits into the *in-degree* and *out-degree* for edges pointing in and out of node i respectively

$$k_i^{out} = \sum_j A_{ij}, \qquad k_i^{in} = \sum_j A_{ij}^T, \tag{2.5}$$

where $\mathbf{A}^T$ is the transposed adjacency matrix. The total number of edges in this case is then $E = \sum_i k_i^{out} = \sum_i k_i^{in} = \sum_{i,j} A_{ij}$. The average degree of a directed graph is $\langle k^{out} \rangle = \langle k^{in} \rangle = \sum_i k_i^{in}/N = E/N$. The weighted degree for undirected graphs, also called *strength*, is the generalization of (2.4)

$$s_i = \sum_j W_{ij}, \tag{2.6}$$

where $W_{ij} \geq 0$ is the weighted adjacency matrix.

More interesting non-local measures include the *betweenness*, *closeness* and *eigenvector* centrality. The betweenness centrality $\mathbf{c}^B$ is a measure of the amount of information transmitted by shortest paths along each node. It is defined as the fraction of shortest paths between all pairs of nodes that also pass through each node

$$c_i^B = \sum_{k,l \neq i} \frac{\mathcal{N}_{kl}(i)}{\mathcal{N}_{kl}}, \tag{2.7}$$

[3] Some would argue that this is not really a centrality as it is a local property of the node, contrary to all other centralities.

[4] We will often relax the discreteness of the variable k and consider instead of $P(k)$, the probability density function [103] $\mathcal{P}(k)$ as the distribution for the continuous variable, so that $\mathcal{P}(k)dk$ gives the probability that a randomly chosen node in the graph has degree with value in the interval $(k, k+dk)$.

where $\mathcal{N}_{kl}$ is the total number of shortest paths from node k to node l and $\mathcal{N}_{kl}(i)$ is the number of these paths that also pass through node i. The closeness centrality $\mathbf{c}^C$ is the reciprocal average distance from a given node to the rest of the graph

$$c_i^C = \frac{1}{\sum_{j \neq i} D_{ij}}. \tag{2.8}$$

The eigenvector centrality $\mathbf{c}^E$ is defined by the components of the leading eigenvector of the adjacency matrix

$$c_i^E = \lambda_{max}^{-1} \sum_j A_{ij} c_j^E, \tag{2.9}$$

where λ_{max} is the largest eigenvalue of $\mathbf{A}$.

Finally the k-core centrality [247, 94] is obtained from the maximal connected subgraph composed of nodes that have at least k neighbors within the set itself. The k-core (or k-shell) decomposition of the graph is the iterative procedure that classifies all nodes in shells of increasing connectivity. Each node is labeled with an integer (the k-core index k_c) that equals the largest k value of k-cores to which the node belongs. The k-shell is the set of nodes that are part of the k-core but not part of the $(k+1)$-core. Then we say that a node has *coreness* k if it belongs to the k-shell. The number of cores using this decomposition gives the *degeneracy* of the graph [99].

2.1.3. Network models

The functional form of the statistical distributions characterizing large-scale networks defines two broad network classes. The first refers to the so-called statistically homogeneous networks. In this case the distributions characterizing the degree and other properties such as nodes betweenness or edges weights have functional forms with fast decaying or "light" tails such as the Poisson distribution. The limiting case of this class is the regular lattice (RL) in d spatial dimensions, defined in the thermodynamic limit by $\mathbb{Z}^d$. In this case each node has the same number of neighbors $k_i = \langle k \rangle = 2d$, $\forall i$ and the degree distribution $\mathcal{P}(k)$ reduces to a Dirac delta centered at $\langle k \rangle$. The second class concerns networks with statistically heterogeneous connectivity and weight patterns usually corresponding to skewed and heavy-tailed distributions. We analyze both classes by defining the three major models of random networks. The first two models, the Erdös-Rényi and Watts-Strogatz model, belong to the class of homogeneous or Poisson random graphs, with very small fluctuations of the degree. Finally we discuss the Barabási-Albert model as the reference model for heterogeneous graphs that go beyond the Poissonian topology with connectivity fluctuations on virtually all scales.

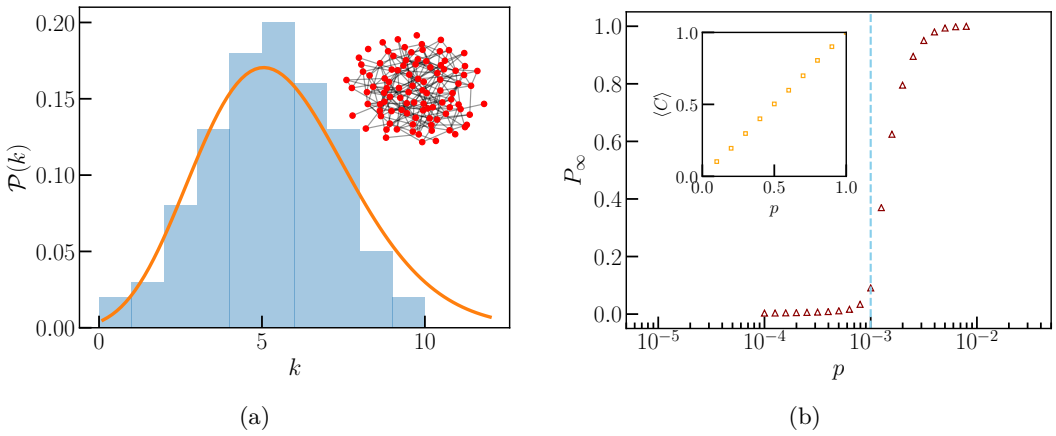


Figure 2.1: (a) Degree distribution for an ER graph (top right) with $N = 100$ and edge creation probability $p = 0.05$, with the best curve fitting of a Poisson distribution with mean equal to the average degree of the graph $\langle k \rangle \approx 5$. (b) The probability P_∞ that a node belongs to the largest connected component of an ER graph with $N = 1000$ nodes as a function of the edge creation probability p. The vertical dashed line identifies the critical probability $p_c = 1/N$. In the inset the average clustering $\langle C \rangle$ as a function of p.

Erdös-Rényi (ER) model

The study of random graphs was started with the discovery of the transition from a disconnected graph to the birth of a giant component, that corresponds to the percolation threshold in condensed matter [52]. The simplest model, introduced by Erdös and Rényi in 1959 [100], is based on the idea that all the edges have the same probability of existence. The graph is built assuming no knowledge of the principles that guide and characterize the creation of connections of real networks. Given N nodes we draw an edge with a certain probability p. This corresponds to sampling every one of the possible $E_{max} = N(N-1)/2$ edges and drawing the edge with connection probability p. The expectation value of the size of the graph is then $\langle E \rangle = pN(N-1)/2$ and the average degree is simply

$$\langle k \rangle = 2\frac{\langle E \rangle}{N} \approx pN. \tag{2.10}$$

The probability to have a node whose degree is k is composed of k times a successful event whose probability is p and $(N-1-k)$ times an unsuccessful event whose probability is $(1-p)$. Then we obtain the binomial distribution which can be approximated by a

Poisson distribution in the thermodynamic limit $N \to \infty$ for fixed $\langle k \rangle = pN$, i.e.

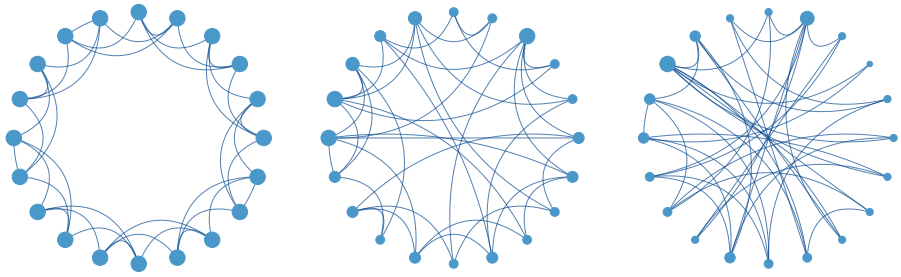
$$\mathcal{P}(k) = \frac{(N-1)!}{(N-1-k)!k!} p^k (1-p)^{(N-1-k)} \approx \frac{\langle k \rangle^k}{k!} e^{-\langle k \rangle}. \tag{2.11}$$

The ER model describes a network where we have two distinct mutually exclusive outcomes for each edge and as a consequence the degree distribution decays exponentially for large k, see Figure 2.1 (a). Crucially, this restricts the allowed degree fluctuations that are predominant in observed real networks [54]. In particular, the second moment of the degree distribution is $\langle k^2 \rangle = \langle k \rangle^2 + \langle k \rangle$.

The diameter of the ER graph can be estimated by considering the number $m_i^{(D)}$ of D-first neighbors of node i. The number of first neighbors of i is simply its degree k_i and if we approximate k_i with its average, then $m_i^{(1)} = \langle k \rangle$. Analogously for the number of second-first neighbors, assuming that none of these are also first neighbors of i, we get $m_i^{(2)} \approx m_i^{(1)} \langle k \rangle = \langle k \rangle^2$. In this approximation the general D-first neighbors of i are simply given by $m_i^{(D)} \approx \langle k \rangle^D$ and since the total number of nodes at leading order is $N = \sum_{k=1}^{D} m_i^{(k)} \approx m_i^{(D)} \approx \langle k \rangle^D$, by taking the logarithm we obtain $D \approx \ln N / \ln \langle k \rangle$. This fact is a common feature, at least qualitatively, of most if not all graph models and explains why the diameter of real networks can remain very small when the number of nodes increases, also known as the *small world* property. Quantitatively, since the network diameter D is often dominated by a few extreme paths, the average shortest path distance $\langle D \rangle$ between node pairs is used in place of D, in order to suppress such fluctuations. Hence a typical formulation of the small world property is

$$\langle D \rangle \approx \frac{\ln N}{\ln \langle k \rangle}. \tag{2.12}$$

In the ER model the simple criteria that a connected component appears when the mean number of second nearest neighbors $m_i^{(2)} = \langle k \rangle^2$ of a randomly chosen node i exceeds the mean number of nearest neighbors $m_i^{(1)} = \langle k \rangle$ gives the condition $\langle k \rangle > 1$ which defines the critical probability $p_c = 1/N$ [93]. The probability P_∞ that a node belongs to the connected giant component of the graph (i.e. to the infinite cluster in the thermodynamic limit) is shown in Figure 2.1 (b). Finally, the clustering coefficient of the ER model is found by considering that for each node the probability that two of its neighbors are connected is the same probability that any other two nodes will be connected and is equal to p. Then for each node i there is a triangle formed with two neighbors with probability p and averaging over all nodes yields $\langle C \rangle = p = \langle k \rangle / N$, see inset in Figure 2.1 (b). In the thermodynamic limit, at fixed average degree, the ER graph has vanishing global clustering. This limitation can be taken care of by introducing an additional mechanism that regulates the position of the edges, as explained next.



$p = 0$ $p = 0.5$ $p = 1$

Figure 2.2: The Watts-Strogatz model for $N = 20$ nodes with $\langle k \rangle = 4$, with node size proportional to its degree. Starting from a regular lattice ($p = 0$), with probability p each link is rewired to a randomly chosen node. The three panels correspond to the regular lattice (left), small-world (center) and random configuration (right), respectively. In the latter all edges have been rewired, so that we recover a Poissonian random graph. Contrary to the ER model, for values of p smaller than unity the graph maintains the high clustering found in regular lattices but in addition the random long-range edges can drastically decrease the distance between nodes.

Watts-Strogatz (WS) model

One of the main drawbacks of the ER model is the absence of clustering in the limit $N \to \infty$, as determined by the exponentially decaying degree distribution. The observed large clustering in real networks has motivated the introduction of a specific mechanism in the graph creation in order to tune the global clustering $\langle C \rangle$ to a desired value. The basic idea is to start with a regular lattice with high $\langle C \rangle$ and gradually rewire each connection with probability p to obtain the desired effect as a function of the induced randomness. The model starts with N nodes disposed on a ring and each of them is symmetrically connected to its $2m$ neighbors. Then each edge is rewired with probability p, from any given node to a randomly selected node by avoiding self-loop formation. This procedure creates shortcuts in the ordered topology of the first graph by keeping constant the number of edges, see Figure 2.2. The average degree is therefore fixed to $\langle k \rangle = 2m$. For the degree distribution one finds for $p \neq 0$

$$\mathcal{P}(k) = \sum_{n=0}^{\min(k-m,m)} \frac{m!}{(m-n)!n!} p^{m-n}(1-p)^n \frac{(pm)^{k-m-n}}{(k-m-n)!} e^{-pm}, \qquad k \geq m. \tag{2.13}$$

In the trivial case $p = 0$ one has instead $\mathcal{P}(k) = \delta(k - \langle k \rangle)$, where $\delta(x)$ is the Dirac delta, since each node has the same number of neighbors as for the d-dimensional regular lattice. Although the WS model is not locally equivalent to the ER graph even in the limit $p \to 1$, the degree distribution in this case reduces to the Poisson distribution found

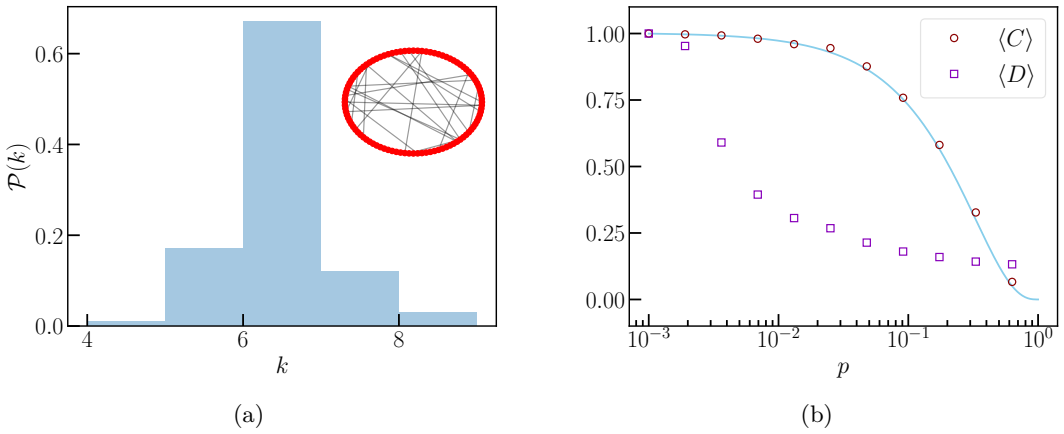


Figure 2.3: (a) Degree distribution for a WS graph (top right) consisting of $N = 100$ nodes, with rewiring probability $p = 0.1$ and number of connected neighbors $m = 3$, yielding an average degree $\langle k \rangle = 6$. (b) Average clustering $\langle C \rangle$ (dark-red circles) with the analytical estimation (2.14) (solid light-blue line) and average shortest-path length $\langle D \rangle$ (violet squares) as a function of the edge rewiring probability p for $N = 1000$ nodes with $\langle k \rangle = 10$. Both quantities are normalized by the respective values at $p = 0$.

for ER and the two models are statistically equivalent. However, contrary to the ER model for the clustering coefficient a small value of randomness $p \ll 1$ yields [24]

$$\langle C(p) \rangle \approx \frac{3}{2} \frac{(m-1)}{(2m-1)} (1-p)^3. \tag{2.14}$$

This expression accounts for the the high global clustering found in real networks, while keeping a small average shortest path length as in the ER model, see Figure 2.3.

In summary, the WS model introduces randomness in a substantially different way from the ER model by allowing for long-range connections starting from a regular topology. This makes it possible to maintain a high clustering while also reducing the distance between nodes. Although the WS model can rarely be used to model real networks, it is a fundamental building block of the theory of complex networks and is the first successful attempt to reconcile an interpolate between the high clustering of nodes with the characteristic small-world effect. However, as for the ER model the degree distribution of the WS model is still far from the fat-tailed heterogeneous distributions characterizing many real networks.

In the next section we will see how the introduction of a new simple ingredient, namely preferential attachment, leads to a completely new paradigm of random graphs that are able to correctly describe many networks observed in Nature.

Barabási-Albert (BA) model

Many real networks display a common statistical property, that is the probability distribution for the degree follows rather than the Poissonian profile a power law

$$\mathcal{P}(k) \sim k^{-\gamma}, \qquad \gamma > 0. \tag{2.15}$$

This implies that different networks have a common qualitative *universality* in their behavior irrespective of their type or the specific mechanism that leads to their creation and evolution, see Figure 2.4. In analogy with the theory of critical phenomena [36], such networks are commonly referred to as *scale-free networks*, because of their lacking of a definite characteristic length scale [54]. Because of this interesting property scale-free networks can be thought of as systems that self-organize into a kind of "critical state". It is important to stress that the recurrent criticism found in the literature that focuses on the poor evidence for power-law behavior in real networks, e.g. see [50], although in principle technically true as no real finite systems can rigorously satisfy (2.15), is conceptually wrong. As only in the thermodynamic limit a phase transition can occur, in strict mathematical terms most of the claimed power laws are actually just an approximation of the actual data that inevitably exhibits upper and lower cutoffs. The power law (2.15) should then be viewed simply as a conceptual modeling framework, just as the Ising model [209] cannot be used to describe real magnets. The correct normalization constant of the distribution implies $\mathcal{P}(k) = Ak^{-\gamma}$ with $A(\gamma) = 1/\int_{k_{min}}^{k_{max}} dk\ k^{-\gamma}$, where k_{max} and k_{min} are the upper and lower cutoff respectively. For the lower cutoff to avoid trivial divergences we always restrict to the connected giant component of the graphs so that $k_{min} \geq 1$. Then the only divergence left is "ultraviolet", that is encountered when we take the thermodynamic limit as $k_{max} \to \infty$. Thus, for $\mathcal{P}(k)$ to be integrable, $A(\gamma)$ must be finite and therefore the exponent must satisfy $\gamma > 1$ for an infinite system. For the average degree we have $\langle k \rangle = A(\gamma) \int_{k_{min}}^{k_{max}} dk\ k^{1-\gamma}$, and using $k_{min} = 1$ yields

$$\langle k \rangle = \frac{(1-\gamma)}{(2-\gamma)} \frac{k_{max}^{2-\gamma} - 1}{k_{max}^{1-\gamma} - 1}. \tag{2.16}$$

Then $\langle k \rangle$ is finite and tends to the constant value $(1-\gamma)/(2-\gamma)$ as k_{max} grows and independently on the cutoff as long as $\gamma > 2$, while it grows as $k_{max}^{2-\gamma}$ for $\gamma < 2$ or logarithmically at exactly $\gamma = 2$. Analogously for the second moment $\langle k^2 \rangle$ we find

$$\langle k^2 \rangle = \frac{(1-\gamma)}{(3-\gamma)} \frac{k_{max}^{3-\gamma} - 1}{k_{max}^{1-\gamma} - 1}, \tag{2.17}$$

which remains finite whenever $\gamma > 3$ and diverges otherwise as the cutoff k_{max} increases. The previous two results combined show that there is a narrow range for the exponent that is satisfied by most real networks [4], namely $2 < \gamma \leq 3$, where even if it is possible

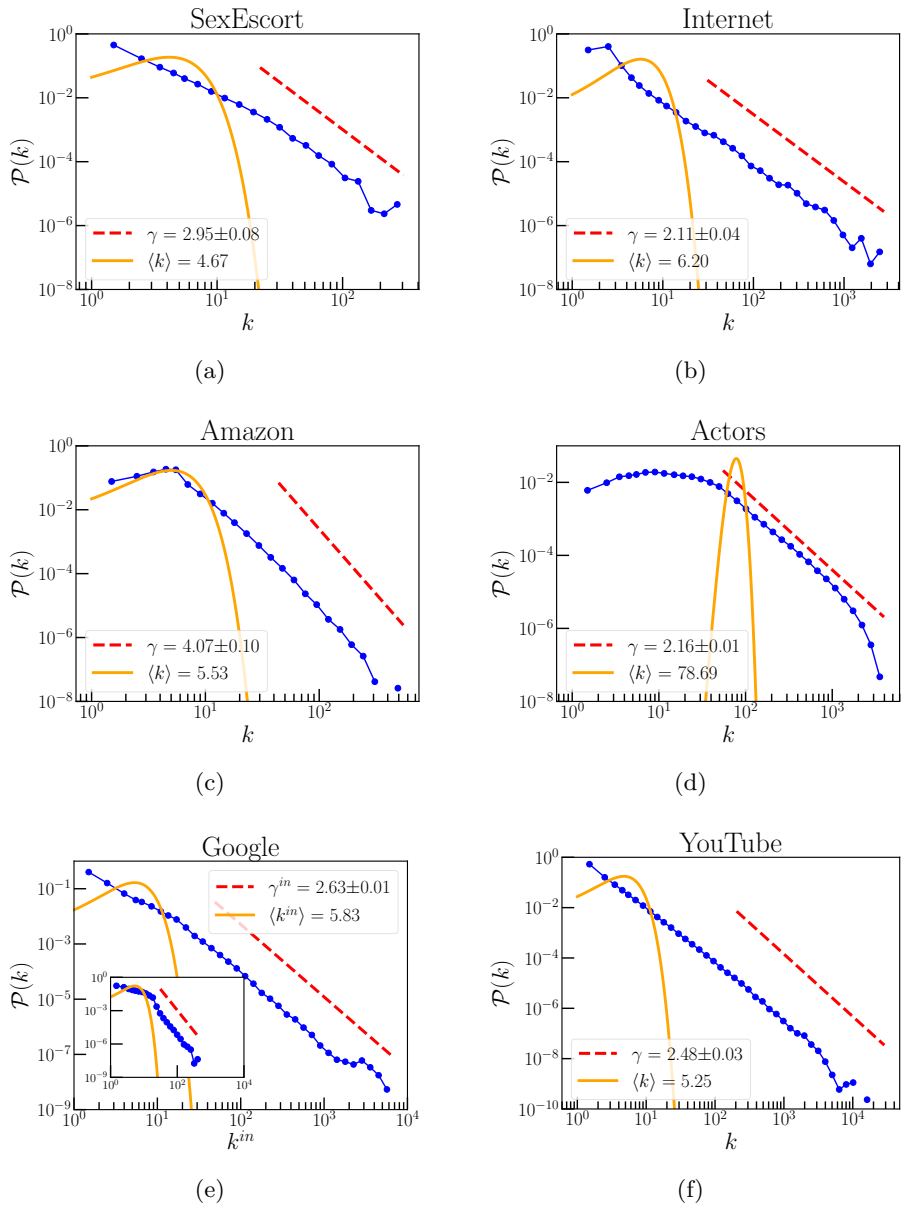


Figure 2.4: Degree distribution (blue circles) of several real networks, from top left: (a) sex buyers and their escorts ($N = 26836$) [234], (b) connections between autonomous systems of the Internet ($N = 34761$) [282], (c) Amazon co-purchases ($N = 334863$) [278], (d) co-appearances of movie actors ($N = 382219$) [18], (e) Google in-hyperlinks and out-hyperlinks (inset) of the directed Web ($N = 875713$) [177], (f) social network of Youtube users and their connections ($N = 1138499$) [196]. In the legend the values of the power-law exponent γ of the degree distribution obtained from the numerical fit (red dashed line) using the method described in [69] and the average degree for the corresponding Poissonian profile (orange solid line).

to define a finite average degree the standard error for this value is of the order of magnitude of the size of the system $k_{max} \sim \mathcal{O}(N)$. The scale-free property of the degree distribution $\mathcal{P}(k)$ implies that each node has a statistically significant probability of having a very large number of connections compared to the average connectivity $\langle k \rangle$ as well as the implicit divergence of $\langle k^2 \rangle$. This fact has profound implications for dynamical processes unfolding on scale-free networks, see Section 2.3.3.

The BA model demonstrates how the combination of two simple ingredients is sufficient to reproduce the universal property (2.15) of scale-free networks found in Nature. These are (i) *growth*, for which new nodes enter the network at some rate, and (ii) *preferential attachment* which forces the newcomers to establish connections preferentially with nodes that already have a large degree [18]. The BA model shows that precisely this "rich-get-richer" philosophy is at the core of the highly heterogeneous topology of observed real networks such as the Internet, the World Wide Web (WWW) [5] or the network representing the large-scale structure of spacetime [168]. The network is built at initial time $t_0 = 0$ starting with a disconnected set of $N_0 = N(t_0)$ nodes with no edges present. Then a new node enters the system at each time step and for each of them m new edges are drawn. The m (per time step) new edges connect newcomer nodes with the old ones extracted with a probability $\Pi(k_i(t))$ proportional to their degree at that moment

$$\Pi(k_i(t)) = \frac{k_i(t)}{\sum_j k_j(t)}. \tag{2.18}$$

Since at every time step only one node is added, the order $N(t)$ and size $E(t)$ of the network at time t is given respectively by

$$\begin{cases} N(t) = N_0 + t \\ E(t) = mt \end{cases} \tag{2.19}$$

This simple update rule is sufficient to naturally produce scale-free distributions for the connectivity given by (2.15). To compute the degree distribution it is convenient to promote k_i to a continuous function of time in order to take its derivative with respect to t so that the degree probability distribution $P(k_i)$ is equivalent to the probability density function. Since the probability to add an edge to i is proportional to the probability (2.18) and the change of connectivity in one time step is m, the total derivative of the degree is

$$\dot{k}_i(t) = m\Pi(k_i(t)) = \frac{k_i(t)}{2t}, \tag{2.20}$$

where we have used $E(t) = \sum_j k_j(t)/2$ and the second of (2.19). By separating the variables we can perform the integration with the initial condition that node i was

added at time t_i with connectivity $k_i(t_i) = m$, yielding the solution

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\beta}, \qquad \beta = \frac{1}{2}. \tag{2.21}$$

This shows that the degree of each node grows as the square root of time and is inversely proportional to its insert time t_i, i.e. oldest nodes will have higher degree as time is increased. By squaring the previous relation, the probability that a node i has degree lower than k is

$$P(k_i(t) < k) = 1 - P\left(t_i \leq \frac{m^{1/\beta}t}{k^{1/\beta}}\right). \tag{2.22}$$

By noting that nodes enter at a constant rate we conclude that their distribution is uniform in time, i.e. $\mathcal{P}(t_i) = A$, where the constant A is defined by imposing the normalization $1 = A\int_0^{N(t)} dN$. Using the first of (2.19) gives $\mathcal{P}(t_i) = 1/(N_0 + t)$ and since for a uniform distribution $P(t_i \leq a) = \int_0^a d\xi\ \mathcal{P}(\xi) = a\mathcal{P}(t_i)$, the cumulative probability function becomes

$$P(k_i(t) < k) = 1 - \frac{m^{1/\beta}t}{N_0 + t}k^{-1/\beta}. \tag{2.23}$$

Taking the derivative of the cumulative probability function we finally obtain the probability density function

$$\mathcal{P}(k) = \frac{d}{dk}P(k_i(t) < k) = \frac{m^{1/\beta}t}{\beta(N_0 + t)}k^{-\gamma}, \qquad \gamma = \frac{1}{\beta} + 1. \tag{2.24}$$

Thus, contrary to the ER and WS random networks, the degree distribution of the BA model is a pure power law 2.15 with exponent $\gamma = 3$. Interestingly, the exponent is independent of the parameters m and N_0, which is a trademark of the universality found in structures driven by growth and preferential attachment. Different values of the exponent in the range $\gamma \geq 2$ can be obtained by considering a preferential-attachment probability $\Pi(k_i(t)) \sim (C + k_i(t))$, which yields $\gamma = 2 + C/m$ where C is a constant [95]. A similar result was also obtained using a general ranking measure, opposed to the basic degree centrality [108]. Generalizations of the preferential attachment rule have been considered to allow the various nodes to have an intrinsic ability or *fitness*, to compete for edges at the expense of other nodes, yielding for the degree distribution a power law with logarithmic corrections [34]. In general random graphs can be extended in a variety of ways to make them more realistic. In addition to the BA model, a simple way to incorporate non-Poisson degree distributions observed in Nature is to consider the configuration model [31, 33]. The latter generates a network starting from any

given degree sequence sampled from a chosen degree distribution $\mathcal{P}(k)$, such as (2.15). For scale-free networks the small world effect (2.12) found in random graphs with no preferential-attachment, becomes even stronger. This *ultra-small world* effect [70] is quantified, for values of the degree distribution exponent $\gamma \in (2, 3)$, by a diameter scaling as $\langle D \rangle \approx \ln \ln N$. Note however that the BA model falls into the marginal case with exponent $\gamma = 3$ where the ultra-small world property does not hold and a similar version of the previous small-world effect is recovered, namely $\langle D \rangle \approx \ln N / \ln \ln N$. Finally, the clustering the BA model is further increased compared to random graph models of same size to $\langle C \rangle \sim (\ln N)^2 / N$ [17].

In summary, the BA model is the benchmark for *complex networks*, just as the Ising model is a reference model for (equilibrium) second order phase transitions [36]. Conveniently, we can use the BA model to answer the question: what actually is a complex network? A possible answer is a (generally large) network with non-trivial topological properties. Contrary to the expected regular and Poissonian topologies described by regular lattices and random networks, the large heterogeneity and heavy-tail properties appear to be common characteristics of a large number of real-world networks, along with other complex topological features, such as hierarchies and communities [123, 229]. The great amount of evidence that scale-free topologies are ubiquitous and emerge as the natural outcome of the evolution of networks can be hardly considered as incidental. This universal behavior is the trademark of general organizing principles that are at the core of the evolution of natural systems in very different contexts.

2.2. Random walks and diffusion on networks

Random walks [218] are one of (if not) the simplest stochastic process that one can define. They are widely used in a variety of applications ranging from physics, economy, mathematical ecology, biology, psychology and even sport statistics [275, 87, 3, 20, 38, 270, 127, 28, 68, 118]. Here, we focus mainly on the category of random walks related to *Markov chains* [206], i.e. in discrete time and discrete space, and in particular on their properties on networks.

Random walks are often used as a model for diffusion. For this reason great effort has been put into understanding the impact of network architecture on random-walk dynamics. The navigation and exploration of complex networks are obviously affected by the underlying connectivity and represent a challenge with many practical applications. To name a few, information discovery and retrieval rely precisely on understanding the properties of random walks and diffusion phenomena in complex networks.

A random walk in discrete time $n \geq 0$ on the lattice $\mathbb{Z}^d$ in d dimensions is described by a sequence of independent, identically distributed random variables $\{X_n\}_{n \geq 0}$ and by the probability density $p_x(n)$ that the walker is located at a point $x \in \mathbb{Z}^d$ after n discrete steps. The process is described by the sum $S_n = \langle x \rangle + \sum_i^n X_i$, where $\langle x \rangle \in \mathbb{Z}^d$

is the initial position, and $p_x(n)$ can be considered as the distribution of the variable $S_n/\sqrt{n}$. This is the simplest mathematical model of linear Brownian motion [170] and also one of the simplest examples of a Markov chain [206]. For simple (unbiased) random walks the process is symmetric: at each time step the walker makes a jump in one of the $2d$ possible directions drawn form the same probability distribution p_x, e.g. on the line ($d = 1$) we have[5] $p_x = (\delta_{x+1} + \delta_{x-1})/2$. If one relaxes the assumption of discrete space to the real numbers $\mathbb{R}^d$ then more interesting distributions can be considered, e.g. Gaussian as shown in Figure 2.5 (a), decreasing exponentials or power laws.

If the first two moments of p_x are finite, since the jumps are independent in each direction, there are no cross-correlation terms and the solution converges asymptotically (for large n) to the Gaussian distribution

$$p_x(n) = \frac{\exp\left(-x^2/4\mathcal{D}n\right)}{(4\pi\mathcal{D}n)^{d/2}}, \tag{2.26}$$

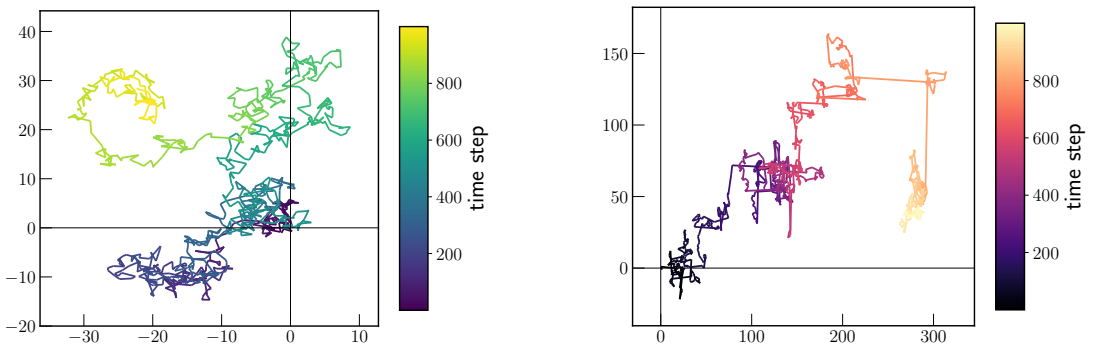
where $\mathcal{D} = \langle x^2\rangle/(2dn)$ is the *diffusion coefficient* and we have assumed that the walker started at the origin so that $\langle x\rangle = 0$. For simple random walks in $\mathbb{Z}^d$, Pólya's theorem [227, 207] states that the trajectory fills the plane densely (the random walk is *recurrent*) for $d \leq 2$ while large regions of space are never visited if $d > 2$ (the random walk is *transient*) [96]. A simple computation reveals that a random walk is also a discrete *fractal* [187]. The fundamental length scale ξ for n-steps random walks, with finite second moment of p_x, scales as $\xi \sim n^{1/2}$. Instead, the "volume" occupied by the walk is measured by the number of steps n. Thus the "mass" of the walk scales with its length as $n \sim \xi^{d_f}$, where $d_f = 2$ is the *fractal dimension* of the random walk [52, 146].

The asymptotic result (2.26) obtained for discrete time steps is a direct application of the *central limit theorem* [103] for the sum of the sizes of the moves, which are independent random variables[6]. This asymptotic regime is well-defined because the underlying space (e.g. the integers $\mathbb{Z}^d$) is infinitely large. In situations in which the second moment of p_x diverges, the process exhibits *superdiffusion* and the probability profile deviates from the Gaussian profile. A generalized central limit theorem [126] for distributions p_x with infinite variance shows that the probability profile in this case converges to a symmetric (α-stable) Lévy distribution [163]. Although an explicit analytical expression for such distributions is not known in the general case, their characteristic function, i.e.

[5] This generalizes in arbitrary dimension to

$$p_x = \frac{1}{2d}\sum_{i}^{d}\left(\delta_{x_i+1} + \delta_{x_i-1}\right)\prod_{k\neq i}^{d}\delta_{x_k}. \tag{2.25}$$

[6] The same result applies also to Gaussian distributed jumps and in general for *any* distribution of single independent jumps as long as the first and second moments are defined.



(a) (b)

Figure 2.5: Random walks in $\mathbb{Z}^2$ over 10^3 time steps with (a) Gaussian jumps centered at the origin with unitary variance converging to ordinary Brownian motion and (b) Lévy flights $p_x \sim |x|^{-1-\alpha}$ with exponent $\alpha = 1.5$.

their Fourier transform, has the form

$$\widetilde{p}(k, n) \sim \exp(-a|k|^{\alpha}), \tag{2.27}$$

where a and $\alpha \in (0, 2)$ are the *scale parameter* and the *distribution index*, respectively. As the Fourier transform of the Gaussian profile (2.26) is again a Gaussian $\widetilde{p}(k, n) \sim \exp(-a|k|^2)$, it can be considered as the limiting case of a stable distribution with index $\alpha = 2$ and finite variance[7].

A particularly interesting case of scale-free random walks that do not possess a second moment are the so-called *Lévy flights* [162] with moves drawn from the power-law distribution $p_x \sim |x|^{-1-\alpha}$, where $\alpha \in (0, 2)$ is the Lévy exponent, which is also the index of the associated stable distribution. Superdiffusive random motion is widely known in Nature and it is found in a variety of physical and biological systems, ranging from transport in chaotic systems, human mobility and money circulation, the foraging behavior of bacteria to hopping processes along a polymer [122, 48, 164].

Interestingly, the Lévy distribution converges asymptotically to the profile of the Lévy flight, i.e. $p_x(n) \sim |x|^{-1-\alpha}$ for $x \to \pm\infty$ [164]. Due to the divergence of the second moment, Lévy flights are characterized by the presence of extremely long jumps and self-similar trajectories, with fractal dimension $d_f = \alpha$ [52, 146], on all scales with clusters of shorter jumps intersected by long ones. A realization of a Lévy flight is shown in Figure 2.5 (b). As it can be seen by comparing with Figure 2.5 (a) the lack of a finite second moment is reflected by a motion over many different scales. In Figure 2.6 we show the distance from the origin of two random walks with Gaussian steps (lower

[7]The case with index $\alpha = 1$ corresponds instead to the Cauchy-Lorentz distribution.

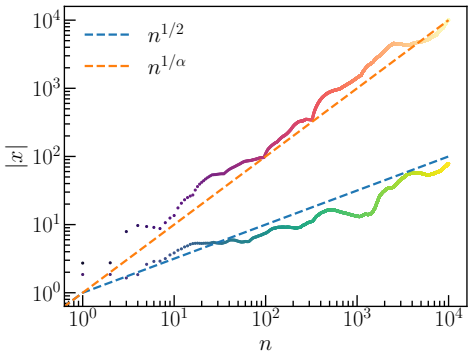


Figure 2.6: Distance from the origin for an ordinary random walk with Gaussian steps (lower trajectory) and Lévy flights with distribution index $\alpha = 1$ (upper trajectory) as a function of the time step n with color changing from dark to light accordingly (color maps as in Figure 2.5). The dashed lines indicate the asymptotic scaling in the respective cases. Clearly, the random walk with Lévy flights is superdiffusive with distance from the origin following asymptotically the power law $|x| \sim n^{1/\alpha}$.

trajectory) and Lévy flights (upper trajectory), respectively. For the former we have ordinary diffusive behavior with distance scaling as $|x| \sim n^{1/2}$ with the number of steps n, while the latter displays superdiffusive behavior $|x| \sim n^{1/\alpha}$ with $\alpha < 2$.

Given a graph $G(\mathcal{V}, \mathcal{L})$ with adjacency matrix A_{ij}, the countable set $\mathcal{V}$ defines the possible states of a Markov chain [206] with *transition matrix* given by the conditional probability

$$P_{ij} = P(X_{n+1} = j | X_n = i) = \frac{A_{ij}}{\sum_l A_{il}}. \tag{2.28}$$

By construction P_{ij} is row stochastic, i.e. $\sum_j P_{ij} = 1$, and so is every power P_{ij}^n which defines the probability to jump from i to j after n time steps. Being independent on the time step n, the transition probability matrix is the *propagator* of a stationary process defined by the discrete time random walk $\{X_n\}_{n\geq 0}$. We always consider *irreducible* chains, also called *ergodic*, where the graph is composed of a single communicating class. We can use the transition matrix to evolve the probability density $p_k(n)$ for the walker to be located at node k as

$$p_i(n+1) = \sum_k p_k(n) P_{ki}. \tag{2.29}$$

The asymptotic limit of the previous equation yields the (invariant) stationary density of the chain defined by

$$\pi_i = \lim_{n\to\infty} p_i(n), \tag{2.30}$$

which satisfies $\pi_i = \sum_k \pi_k P_{ki} = \lim_{n\to\infty} P_{ki}^n$, as it can be checked by iteration of (2.29). Therefore, the stationary density π_i and the vector of all ones $e_i = 1, \forall i$, are the left and right eigenvectors of $\mathbf{P}$ with eigenvalue 1, respectively. For an ergodic chain with finite state space, π_i is unique and it is proportional to the degree k_i. From the definition

(2.28) undirected networks satisfy the *detailed balance* condition

$$\pi_i P_{ij} = \pi_j P_{ji}, \tag{2.31}$$

where $\pi_i = k_i / \sum_l k_l = k_i/2E$, and the associated Markov chain is *reversible*. As we will see next, the famous web-ranking algorithm PageRank [44] is defined as the stationary density of a modified random-walk with update rule such as (2.29).

Arguably, the most famous centrality measure based on random walks on graphs is *PageRank* [210], originally devised for ranking web pages. PageRank has been used and generalized for numerous applications including ranking of academic journals and papers, professional sports, disease-gene identification, systemic risk in financial networks, ordering of functions in Linux, prediction of traffic flow and human movement, recommendation systems in online marketplaces, image search engines, identifying community structure in networks, and much more [124]. The PageRank vector x_i is defined as the stationary density of a discrete time random walk on a graph that is a modification of the original graph with occupation probability satisfying (2.29), in order to guarantee that the stationary density always exists. The node stationary density of the random walk on the modified graph defines the PageRank score of that node. The modification consists in allowing the walker to "teleport" to other nodes. The evolution equation for the modified random walk reads

$$x_i(n+1) = \alpha \sum_k x_k(n) P_{ki} + (1-\alpha) g_i, \tag{2.32}$$

where $(1-\alpha)$ is the teleportation probability, which gives the probability to randomly relocate during the walk. For practical application one usually assumes that $\alpha = 0.85$ [124]. In the limit $\alpha \to 1$ we indeed recover the Markov chain equation (2.29). The *preference vector* g_i, that satisfies the constraint $\sum_j g_j = 1$, determines the conditional probability that a walker teleports to node i. The standard choice for the preference vector is the uniform distribution $g_i = e_i/N$, where $e_i = 1$ for all i's. Generalizations to non-uniform preference vectors have also been studied in [171]. The stationary solution of (2.32) defines the PageRank score x_i. By construction, PageRank is normalized to unity, i.e. $\sum_j x_j = 1$. Similarly to the eigenvector centrality, defined as the dominant right eigenvector of the adjacency matrix A_{ij}, the PageRank centrality can also be defined in terms of the spectra of the *Google matrix*

$$G_{ij} = \alpha P_{ij} + (1-\alpha) e_i g_j. \tag{2.33}$$

PageRank can be computed as the left eigenvector of $\mathbf{G}$ corresponding to the maximum eigenvalue $\lambda_{max} = 1$, i.e.

$$\mathbf{x}^T \mathbf{G} = \mathbf{x}^T. \tag{2.34}$$

2.2.1. Graph Laplacian

The continuous-time limit of the process, obtained by replacing the discrete time step $n \in \mathbb{N}^+$ with $t \in \mathbb{R}^+$, is described by a *master equation*[8]. First, we expand the time-dependent transition matrix

$$\mathbf{P}(t) = e^{\mathbf{Q}t}, \tag{2.35}$$

for $t \to 0$, where Q_{ij} are the *transition rates*, i.e. the transition probabilities per unit time, that define the infinitesimal generator of the continuous-time *Markov semigroup* [104]. By definition, the forward equation $\dot{\mathbf{P}}(t) = \mathbf{P}(t)\mathbf{Q}$ and the adjoint backward equation $\dot{\mathbf{P}}(t) = \mathbf{Q}\mathbf{P}(t)$ are satisfied, while $\mathbf{P}(0) = \mathbf{I}$. The rate matrix satisfies $\sum_k Q_{ik} = 0$ for all i and the diagonal element changed in sign

$$q_i \equiv -Q_{ii} = \sum_{l \neq i} Q_{il}, \tag{2.36}$$

defines the (non-negative) *exit rate* of node i. The expansion of $\mathbf{P}(\Delta t)$ up to order $\mathcal{O}(\Delta t^2)$ reads

$$[\mathbf{P}(\Delta t)]_{kj} \approx (\mathbf{I} + \mathbf{Q}\Delta t)_{kj} = \begin{cases} Q_{kj}\Delta t & k \neq j \\ 1 - \sum_{l \neq j} Q_{jl}\Delta t & k = j \end{cases} \tag{2.37}$$

Using this result and subtracting $P_{ij}(t)$ from the Chapman-Kolmogorov equation

$$P_{ij}(t + \Delta t) = \sum_k P_{ik}(t) P_{kj}(\Delta t), \tag{2.38}$$

after taking the limit $\Delta t \to 0$ and using (2.29), yields the master equation of the continuous-time random walk on graphs

$$\dot{p}_i(t) = \sum_{k \neq i} p_k(t) Q_{ki} - \sum_{k \neq i} p_i(t) Q_{ik}. \tag{2.39}$$

The two terms on the right hand side correspond to the in-flux (gain) and out-flux (loss) of probability respectively, of node i.

The Gaussian profile (2.26) in continuous space $x \in \mathbb{R}^d$ and continuous time $t \in \mathbb{R}^+$ is also the formal solution of the diffusion equation that describes Brownian motion

$$\partial_t p(x,t) = \mathcal{D}\nabla^2 p(x,t), \tag{2.40}$$

[8] The name stands from the original paper in which it first appears as it was the general equation from which all others could be derived [205].

where $\mathcal{D} = \langle x^2 \rangle / (2dt)$ is the diffusion coefficient and $\nabla^2 = \sum_i^d \partial_{x_i}^2$ is the *Laplace* operator[9]. (2.40) is often interpreted as a continuity equation $\partial_t p(x,t) + \nabla J(x,t) = 0$, where the current is assumed to be proportional to the spatial change in concentration i.e. by Fick's law [106] $J(x,t) = -\mathcal{D}\nabla p(x,t)$.

The diffusion equation (2.40) can easily be generalized to describe diffusion processes on networks [201]. Let us suppose we have some concentration of substance $\rho_i(t)$ on each node i. Then we allow the substance to move along the edges, flowing from one node i to an adjacent one j at a rate $\mathcal{D}$. That is, in a small interval of time dt the amount of substance flowing from i to j is equal to $\mathcal{D}(\rho_j(t) - \rho_i(t))dt$. Then the rate at which $\rho_i(t)$ is changing is given by $\dot{\rho}_i(t) = \mathcal{D}\sum_j A_{ij}\left(\rho_j(t) - \rho_i(t)\right)$. Rearranging, one immediately finds the analogous of the ordinary diffusion equation (2.40) on the discrete space of nodes in the graph

$$\dot{\rho}_i(t) = -\mathcal{D}\sum_k L_{ik}\rho_k(t), \tag{2.41}$$

where

$$L_{ij} = \delta_{ij}k_i - A_{ij} \tag{2.42}$$

is the graph Laplacian that corresponds to the negative of its continuous-space version ∇^2. For this reason the matrix (2.42) has a very central role in the characterization of diffusive processes on networks. It is easy to show that on the infinite lattice $\mathbb{Z}^d$ one has indeed $\mathbf{L} = -\nabla^2$ [54, 172]. Since we are considering undirected networks, $\mathbf{L} = \mathbf{L}^T$ and so all Laplacian eigenvalues $\{\lambda_l\}$ are real and non-negative with $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$, where non-negative property comes from the decomposition of the Laplacian in terms of the incidence matrix of the graph [263]. The eigenvalue $\lambda_1 = 0$ is always an eigenvalue because by construction every row and column of $\mathbf{L}$ sums to zero. Then the vector of ones $\mathbf{e}$ is the corresponding eigenvector and the Laplacian matrix is always singular. Interestingly, the second eigenvalue of the graph Laplacian λ_2, called the algebraic connectivity of the network, is non-zero if and only if the network is connected [263]. The formal solution of (2.41) is obtained by writing the vector $\rho(t)$ as a linear combination of the eigenvectors $\{\mathbf{v}^{(l)}\}$ of the Laplacian

$$\boldsymbol{\rho}(t) = \sum_l \mathbf{v}^{(l)} c_l(t), \qquad \mathbf{L}\mathbf{v}^{(l)} = \lambda_l \mathbf{v}^{(l)}, \tag{2.43}$$

where $c_l(t)$ are the time-dependent coefficients of the expansion that determine the

[9] For Lévy distributions the corresponding diffusion equation is formulated in terms of the fractional Laplacian $\sum_i^d \partial_{|x_i|}^{\alpha}$ [163].

solution $\boldsymbol{\rho}(t)$. The diffusion equation (2.41) then becomes

$$\sum_l \left(\dot{c}_l(t) + \mathcal{D}\lambda_l c_l(t)\right) \mathbf{v}^{(l)} = 0, \tag{2.44}$$

and the solution follows immediately as

$$c_l(t) = c_l(0) e^{-\mathcal{D}\lambda_l t}. \tag{2.45}$$

Another useful quantity for describing diffusion on networks is the rescaled Laplacian

$$\widetilde{\mathbf{L}} = \mathbf{K}^{-1}\mathbf{L} = \mathbf{I} - \mathbf{P}, \tag{2.46}$$

where $K_{ij} = \delta_{ij} k_i$ is the diagonal matrix with the degree vector on the diagonal. By definition, the stationary density of the discrete-time random walk on the graph π_i satisfies $\pi\widetilde{\mathbf{L}} = 0$. Then $\pi\mathbf{K}^{-1}$ is an eigenvector of the Laplacian matrix $\mathbf{L}$ with eigenvalue zero. However for connected networks there is only one null eigenvalue corresponding to the eigenvector proportional to the all-ones vector $\mathbf{e}$ while all other eigenvalues are positive. Hence $\pi\mathbf{K}^{-1} = c\mathbf{e}$, with c constant and one recovers the stationary density in the detailed balance (2.31) as $\pi_i = ck_i$, with $c^{-1} = 2E$.

2.2.2. Hitting times

Given a Markov chain $(X_n)_{n\geq 0}$ over the set of states $\{\mathcal{V}\}$, the hitting time for the state j is defined as the minimum number of steps $n_j = \min_n\{n|X_n = j\}$ necessary to enter j[206]. The probability associated to a finite $n_j = n$, given that the random walker started in state i, is the *hitting time probability*

$$H_{ij}(n) = P(n_j = n|X_0 = i). \tag{2.47}$$

The moments $\langle n^k\rangle$ of the probability distribution $H_{ij}(n)$ can easily be obtained from differentiation of the moment generating function[10]

$$\Phi(\lambda) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \langle n^k\rangle = \langle e^{\lambda n}\rangle \,. \tag{2.48}$$

Analogously the cumulants $\langle n^k\rangle_c$ can be obtained from the logarithm of (2.48), which defines the cumulant generating function

$$\Psi(\lambda) = \ln \Phi(\lambda) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \langle n^k\rangle_c \,. \tag{2.49}$$

[10]For imaginary source $\lambda = i\xi$ this is the discrete characteristic function [170].

The first moment of $H_{ij}(n)$ defines the *mean first passage time* (MFPT) matrix $M_{ij} \equiv \langle n_{ij} \rangle$, that is the expected value [154]

$$M_{ij} = \sum_{n=0}^{\infty} n H_{ij}(n). \tag{2.50}$$

The hitting time probability can be computed recursively by making the target node j an absorbing state [206] , i.e. a single closed communicating class, for $i \neq j$ as $H_{ij}(n) = \sum_{k \neq j} P_{ik} H_{kj}(n)$, while $H_{ij}(n) = \delta_{n,0}$ if $i = j$. It is easy to verify that the MFPT matrix also satisfies the recursive relation for all i and j

$$M_{ij} = 1 + \sum_{k \neq j} P_{ik} M_{kj}. \tag{2.51}$$

In its first step, a random walker moves from node i to node k, which produces the 1 on the right-hand side of (2.51). If $k = j$, then the walk terminates at k, resulting in a first-passage time of 1. Otherwise, we seek the first-passage from node k (with $k \neq j$) to node j. This produces the second term on the right-hand side. The previous relation is also valid when $i = j$ giving the so-called mean recurrence times M_{ii}. In matrix notation, we can write the MFPT matrix (2.51) as

$$\mathbf{M} = \mathbf{E} + \mathbf{P}(\mathbf{M} - \mathbf{R}), \tag{2.52}$$

where $E_{ij} = 1, \forall (i,j)$ is the matrix of ones and $R_{ij} = \delta_{ij} M_{ii}$ is the diagonal matrix whose entries are the recurrence times. By left-multiplying (2.52) for the stationary density $\boldsymbol{\pi}$ and using $\boldsymbol{\pi}\mathbf{E} = \mathbf{e}$, where $\mathbf{e}$ is the vector of all ones, and $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$, we obtain the mean recurrence times $M_{ii} = 1/\pi_i, \forall i$, known as the Kac's formula. Besides iterating (2.51) one can compute the MFPT $(N-1)$-dimensional sub-vector $\mathbf{m}^{(j)}$ for $i \neq j$ obtained by vectorising M_{ij} for fixed target j by removing the jth element. For $i \neq j$ (2.51) becomes $m_i^{(j)} = 1 + \sum_{k \neq j} P_{ik}^{(j)} m_k^{(j)}$, where $\mathbf{P}^{(j)}$ is the $(N-1 \times N-1)$-dimensional transition sub-matrix with jth row and column removed. This can be solved immediately and yields

$$\mathbf{m}^{(j)} = (\mathbf{I}^{(j)} - \mathbf{P}^{(j)})^{-1} \mathbf{e}^{(j)}, \tag{2.53}$$

where $\mathbf{I}^{(j)} = \{\delta_{ij}\}_{N-1}$ is the $(N-1 \times N-1)$-dimensional identity matrix, and $\mathbf{e}^{(j)}$ is the vector of all ones of length $(N-1)$. The matrix $(\mathbf{I}^{(j)} - \mathbf{P}^{(j)})$ on the right-hand side is reminiscent of the rescaled graph Laplacian (2.46) and it is known as *grounded Laplacian*, although it is not a Laplacian matrix. Indeed the properties of the true Laplacian are not satisfied by this matrix since the number of nodes is reduced by one (the jth node) while the number of edges stays the same as node j would still be part of the graph.

Contrary to the rescaled Laplacian (2.46), the rescaled grounded Laplacian

$$\widetilde{\mathbf{L}}^{(j)} = \mathbf{I}^{(j)} - \mathbf{P}^{(j)}, \tag{2.54}$$

is always invertible because for (strongly) connected networks, by removing one row and one column we also remove the singularity of $\widetilde{\mathbf{L}}$. Finally, the MFPT can also be obtained by differentiating once the moment (2.48) or the cumulant generating function (2.49) with respect to λ and evaluated at zero argument, i.e.

$$M_{ij} = \left.\frac{\partial \Phi_{ij}(\lambda)}{\partial \lambda}\right|_{\lambda=0} = \left.\frac{\partial \Psi_{ij}(\lambda)}{\partial \lambda}\right|_{\lambda=0}. \tag{2.55}$$

2.3. Spreading processes

We can describe dynamical processes on networks by assigning a variable $\sigma_i(t)$ to each node characterizing its dynamical state at time t. For small magnets on a graph $G(\mathcal{V}, \mathcal{L})$, the variable $\sigma_i(t)$ represents the *spin* at site i (that can be up or down), while in the case of epidemic spreading it indicates if the individual is healthy, infected or when it is the case if he/she has recovered from the disease. The microscopic state of the whole system is defined by the particular configuration of the network variables at each time given by the set $\sigma(t) = \{\sigma_i(t)\}$. A dynamical process is described by the transition to a different configuration $\sigma \to \sigma'$. The basic dynamical description of the system relies on the master equation approach that is formalized by replacing in 2.39 the nodes with a whole system configuration, i.e.

$$\partial_t p(\sigma, t) = \sum_{\{\sigma'\}} p(\sigma', t) Q(\sigma' \to \sigma) - \sum_{\{\sigma'\}} p(\sigma, t) Q(\sigma \to \sigma'), \tag{2.56}$$

Here, the sum is over all allowed system configurations and $Q(\sigma \to \sigma')$ are the transition rates from one configuration to the other. Solving the master equation allows the calculation of the expectation values of all quantities of interest in the system. Given an observable function of the state of the system $O(\sigma)$, its average value at time t is given by $\langle O \rangle = \sum_\sigma O(\sigma) p(\sigma, t)$, where $\langle \ldots \rangle$ is the phase space average. While in most cases the formal solution of (2.56) is impossible to find, the average $\langle \ldots \rangle$ can be obtained of as an average over different stochastic realizations of the evolution of the system starting with identical initial conditions. A system with a well-defined asymptotic limit $\lim_{t\to\infty} p(\sigma, t) = p(\sigma, \infty)$ is said to be in a (time-independent) *stationary state*. For equilibrium systems an explicit form for the stationary distribution $p(\sigma, \infty) = \bar{p}(\sigma)$ is given by the Boltzmann distribution [144]

$$\bar{p}(\sigma) = \frac{\exp(-\mathcal{H}(\sigma)/k_B T)}{Z}. \tag{2.57}$$

Here, T is the temperature, k_B is the Boltzmann constant, $\mathcal{H}(\sigma)$ is the *Hamiltonian* of configuration σ and the normalization $Z = \sum_\sigma \exp(-\mathcal{H}(\sigma)/k_B T)$ is the *partition function*. The equilibrium distribution (2.57) defines the solution of the master equation for equilibrium systems where the detailed balance condition holds

$$\bar{p}(\sigma)Q(\sigma \to \sigma') = \bar{p}(\sigma')Q(\sigma' \to \sigma). \tag{2.58}$$

As for discrete-time random walks on graphs, this relation states that the net probability current between pairs of configurations is zero when $p = \bar{p}$.

We can repeat all previous steps by considering instead of the configuration σ of the whole system a single node label i. By comparing (2.57) with (2.31) we can immediately identify the Hamiltonian associated to a single node i with $\mathcal{H}(i) = -k_B T \ln(k_i)$ and $Z = 2E$ with the partition function, where E is the number fo edges in the graph. In this analogy a non-equilibrium system is described by a directed graph with $A_{ij} \neq A_{ji}$ for some (i, j), so that $k_i P_{ij} = A_{ij} \neq A_{ji} = k_j P_{ji}$ and detailed balance is not satisfied. A wide range of different systems can be found constantly out of the detailed balance condition. In general many non-equilibrium systems are characterized by the presence of absorbing states. These are configurations that can only be reached but not left, such as a node i in a directed graph for which $k_i^{out} = 0$. In this case we always have a non-zero probability current for some configurations so that the temporal evolution cannot be described by an equilibrium distribution. As we will see next, this is precisely what happens in the case of epidemic spreading. With the exception of Chapter 5, we will always consider the equilibrium description of the underlying networks, i.e. undirected graphs where the detailed balance condition (2.31) for the degree (or for strengths in weighted networks) is satisfied. Importantly, the lack of detailed balance does not imply the absence of a stationary state. Indeed while the detailed balance is sufficient to achieve $\partial_t p(\sigma, t) = 0$, it is not a necessary condition.

2.3.1. Non-equilibrium phase transitions

Almost all systems in Nature are open systems coupled to external reservoir such that the exchange of energy, particles or other conserved quantities between the system and the reservoir leads to currents through the system [170, 286]. Such non-equilibrium effects manifest themselves microscopically with the breaking of detailed balance (2.58). Thus the currents between microstates do not balance and there is in general a non-vanishing flow of probability from one state to another. Unfortunately, the canonical formalism of equilibrium statistical mechanics with microstates probability given by (2.57) does not yet exist for general non-equilibrium systems. In contrast to equilibrium systems, for non-equilibrium systems time is an essential degree of freedom and the relaxation toward an *equilibrium stationary state* may occur only if some ergodicity requirements and detailed balance are satisfied. On the contrary if detailed balance is not satisfied, a

much richer collective behavior over large scales can be found in large complex systems. At equilibrium, the fine tuning of a control parameter (e.g. for a thermodynamic system the temperature T), leads the system to undergo a (second-order) *continuous phase transition* [212, 36]. At the critical point T_c dynamically created long-range correlations emerge, even if the original microscopic interactions are short-ranged. The phase transition is described by an order parameter [212] having a non-zero value in the ordered phase whereas it vanishes in the disordered phase when increasing the control parameter T. The "standard model" for equilibrium (second-order) phase transitions is the Ising model. The order parameter is the average magnetization density $\rho = N^{-1}\sum_i \langle\sigma_i\rangle$, where $\langle\sigma_i\rangle$ is the thermal average of the spin variable $\sigma_i = \pm 1$ over the canonical ensemble (2.57) with Hamiltonian $\mathcal{H} = -J\sum_{ij} A_{ij}\sigma_i\sigma_j$. Here, J is the interaction energy and A_{ij} the adjacency matrix, while the thermal average $\langle\sigma_i\rangle$ coincides with the long-time average if the system is ergotic [119]. Second-order phase transitions are characterized by *universal* long-range correlations that are independent on the microscopic details of the model. For equilibrium systems these universal properties are now been understood in the framework given by the *renormalization group* [152, 276, 56, 253]. At criticality very different systems might exhibit the same behavior, with the same critical exponents and scaling functions, thus belonging to the same *universality class*. Decreasing the control parameter from high temperature (upper left panel in Figure 2.7), exactly at $T = T_c$ a ferromagnetic (ordered) cluster spanning the size of the system first appears (upper central panel in Figure 2.7) and the order parameter increases as T is further decreased (upper right panel in Figure 2.7).

Much of what is known about equilibrium phase transitions can be extended to the non-equilibrium case. Luckily, the central concept of universality, which played a central role at equilibrium, can as well be applied to non-equilibrium systems. A very important class of non-equilibrium phase transitions happens with irreversible microscopic dynamics so that detailed balance is broken and the stationary states cannot be equilibrium states. In this situation, one has a second-order phase transition from a fluctuating ordered state into an *absorbing state* that, by definition, once reached can never be left. As for equilibrium phase transitions, such *absorbing phase transitions* exhibit universal features that in regular lattices are determined only by the space dimensionality, the number of components of the order parameter and the symmetry properties of the system. Any other property is related to irrelevant observables in the sense of the renormalization group. The most important class of absorbing phase transitions is *directed percolation* (from the Latin *percolare* = to filter), originally introduced as a model for directed random connectivity [45], that shares the same type of transition for models describing the spreading of epidemics following essentially the same *reaction-diffusion* scheme. Directed percolation, also known as *contact process* [134], is widely considered as the Ising model of non-equilibrium phase transitions.

Reaction-diffusion processes are sometimes referred to as diffusion-limited reactions [232]. The term diffusion-limited refers to the fact that the reaction itself is fast and

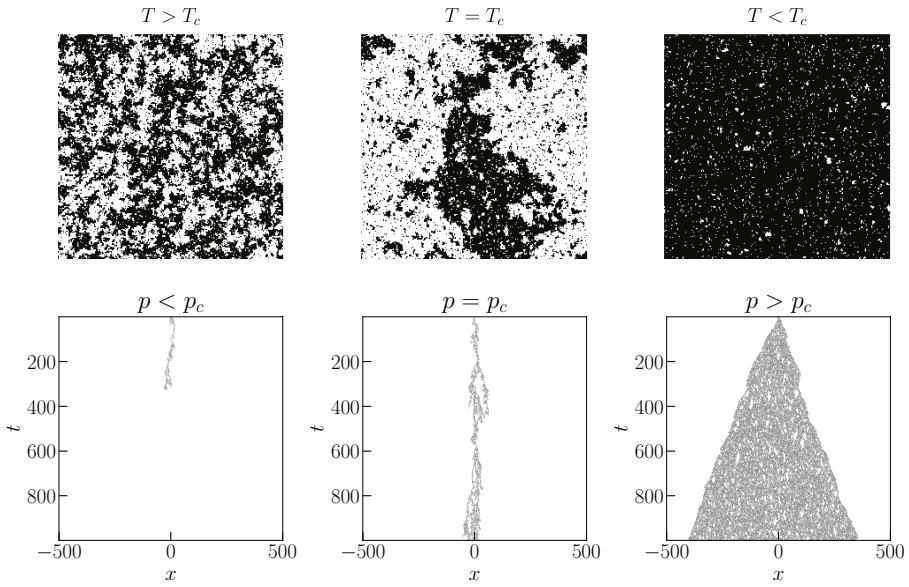


Figure 2.7: Upper panel: From left to right three equilibrium configurations of the Ising model reached after 10^4 MonteCarlo steps of Metropolis dynamics [202, 136] on a two-dimensional lattice with $N = 256^2$ spins above, at and below the critical temperature (in units of the spins interaction energy J) $T_c = 2/\ln(1+\sqrt{2})$ [209]. Lower panel: from left to right three realizations over $t_{max} = 10^3$ time steps (vertical axis) of directed percolation in one dimension (horizontal axis) below, at and above the percolation threshold p_c.

the overall kinetics is controlled by the transport mechanism that brings reactive pairs together. As we discuss in Chapter 4, because the reaction occurs when particles first meet, first-passage theory provides a useful perspective for understanding the kinetics for such processes.

Systems that are governed by an interplay of reaction and diffusion processes are relevant to many problems in physics and other disciplines as diverse as chemical reactions, population evolution, epidemic spreading and many other spatially distributed systems [74]. The classic example is provided by chemical reactions, in which different molecules or atoms diffuse in space and may react whenever in close contact. In fact, many different systems seem to be governed by a set of partial differential equations

$$\partial_t \rho(x,t) = \mathcal{D}\nabla^2 \rho(x,t) + R[\rho(x,t)], \tag{2.59}$$

typically describing the concentration $\rho(x,t)$ of molecules at position x at time t. For a magnetic system, the field $\rho(x,t)$ is the local average magnetization [212] $\rho(x,t) \sim \langle \sigma(t) \rangle_{\mathcal{B}(x,r)} = r^{-d} \sum_{j \in \mathcal{B}(x,r)} \sigma_j(t)$, where the average is taken over the d-dimensional ball $\mathcal{B}(x,r)$ centered in x of radius r. Then, by approximating the discrete space of the set

of vertices $\mathcal{V} = \{i\}$ with continuous positions $x \in \mathbb{R}^d$, the order parameter $\rho(x,t)$ is the relevant variable that describes the critical behavior of the magnetic system instead of discrete field $\sigma_i(t)$.

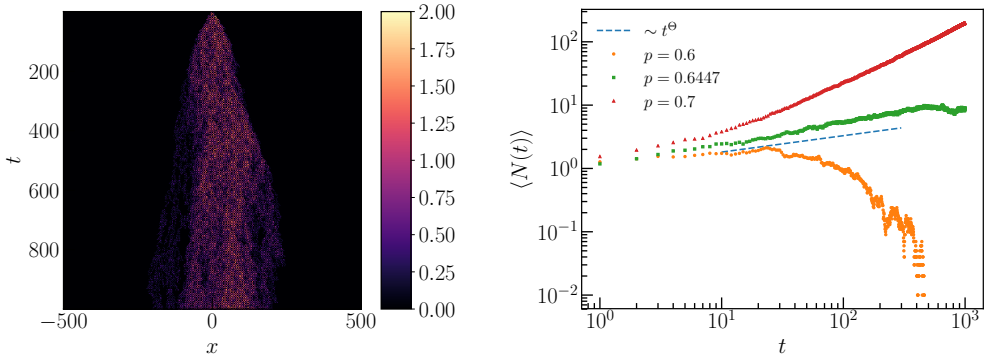
The reaction-diffusion equations (2.59) are obtained simply by adding a reaction term $R[\rho(x,t)]$ to the diffusion equation (2.40). The first results on reaction-diffusion problems using partial differential equations were obtained in 1937 with the studies of autocatalytic reactions $A+B \to B+B$ for two species of particles, by Fischer [107] and by Kolmogorov, Petrovskii, and Piskunov [257]. The particular reaction considered by Fisher corresponds precisely to the simplest susceptible-infected model of epidemic spreading.

Various model for percolation, both directed or isotropic (i.e. undirected), have been introduced. In bond percolation models the sites of a lattice represent the pores of a filter and neighbors pores can be connected with *occupation probability* p, mimicking the irregularities of a network. The percolation probability controls the microscopic connectivity and influences the macroscopic permeability of the filter. As for the Ising model, there exists a critical threshold p_c that marks the onset of a continuous phase transition from an *active phase* ($p > p_c$) to an *absorbing phase* ($p < p_c$). Contrary to isotropic percolation, which can be mapped exactly to the equilibrium Potts model [138], directed percolation is not exactly solvable. The model can be conveniently interpreted as a stochastic many-particle process far from equilibrium, as follows. We label active (wet/spin up) sites as particles A corresponding to the state variable $\sigma(x,t) = 1$ at time t and inactive (dry/spin down) sites as vacancies $\emptyset$, corresponding to $\sigma(x,t) = -1$. Directed percolation is the reaction-diffusion process combining single-particle diffusion with the three reactions: (i) $A \to \emptyset$ (particle removal), (ii) $A \to A + A$ (offspring production) and (iii) $A + A \to A$ (coalescence). In one spatial dimension the process starts with a single particle A at the origin (seed). At each time step with probability p the left site becomes occupied and similarly for the right site. A realization of the process in the three scenarios below, at and above the critical point $p_c \approx 0.6447$ is shown in the lower panel of Figure 2.7. In analogy with equilibrium configurations of the Ising model, increasing the control parameter from low probability, exactly at p_c a cluster spanning the whole system first appears (lower central panel in Figure 2.7). Exactly at criticality the number of particles $\langle N(t) \rangle$, averaged over many realizations of the process, is found to grow asymptotically as a power law with universal exponent Θ, see Figure 2.8.

In order to analyze the off-equilibrium dynamics, it is convenient to set up a continuous description in terms of the coarse-grained probability density (analogous to the magnetization density $\rho(x,t) = \mathcal{P}[\sigma(x,t) = 1]$) which obeys the phenomenological stochastic *Langevin equation* [285, 138]

$$\partial_t \rho(x,t) = -\frac{\delta \mathcal{H}}{\delta \rho(x,t)} + \eta(x,t). \tag{2.60}$$

Here $\eta(x,t)$ is a white noise with variance $\langle \eta(x,t)\eta(x',t') \rangle = \kappa[\rho(x,t)]\delta^d(x-x')\delta(t-t')$,



(a) (b)

Figure 2.8: (a) Average realization of directed percolation over $t_{max} = 10^3$ time steps (vertical axis) and one dimension (horizontal axis) at the percolation threshold p_c, with color scaling according to the average site occupation. (b) Average number of occupied sites $\langle N(t) \rangle$ as a function of time for subcritical (orange), critical (green) and supercritical (red) directed percolation. The scaling asymptotically valid at criticality $\langle N(t) \rangle \sim t^{\Theta}$ from the numerical fit (dashed blue line) yields $\Theta \approx 0.25$ (true value is $\Theta \approx 0.31$).

where $\kappa[\rho(x,t)] = \rho(x,t)$ ensures that the absorbing state $\rho(x,t) = 0$ does not fluctuate and $\mathcal{H}$ is the Ginzburg-Landau functional with cubic interaction[11]

$$\mathcal{H} = \int d^d x \left(\frac{\mathcal{D}}{2} (\nabla \rho(x,t))^2 + \frac{m^2}{2} \rho(x,t)^2 + \frac{g}{3} \rho(x,t)^3 \right). \tag{2.62}$$

The reaction term $R[\rho(x,t)]$ in (2.59) is given by the functional derivative with respect to ρ of the potential in the functional $\mathcal{H}$. In $\mathbb{R}^d$, a dimensional analysis reveals that the noise $\eta(x,t)$ becomes irrelevant in spatial dimensions $d > 4$ [138]. Higher order terms in the order parameter or the noise are also found to be irrelevant under renormalization group arguments [138]. The coarse-grained description for the particle density $\rho(t)$ that neglects any spatial information yields $\dot{\rho}(t) = R[\rho(t)]$, where the reaction for directed percolation

[11] As before here we are assuming that the position x changes smoothly, approximating the node location i with a continuous variable x in $\mathbb{R}^d$. For a ferromagnetic system $\kappa[\rho(x,t)] = 2T$ [138] and the off-equilibrium dynamic is described in terms of the Ginzburg-Landau functional of the ρ^4 theory [212, 285]

$$\mathcal{H} = \int d^d x \left(\frac{\mathcal{D}}{2} (\nabla \rho(x,t))^2 + \frac{m^2}{2} \rho(x,t)^2 + \frac{g}{4} \rho(x,t)^4 \right), \tag{2.61}$$

where $m^2 \sim (T - T_c)$. Because of the (reflection) Z_2 symmetry, the functional is minimized by the two equivalent configurations $\rho(x,t) = \pm\sqrt{-m^2/g}$. In the field theory language, directed percolation is thus described by a ρ^3 theory that is no more invariant under the Ising Z_2 symmetry.

consists of a linear term, combining particle removal and offspring production, plus a quadratic interaction term accounting for coalescence [138], i.e. $R[\rho(t)] = -m^2\rho(t) - g\rho(t)^2$. By taking into account also the spatial dependence that gives rise to diffusion as well as the density fluctuations, but still in a mean-field picture where the occupancy of each site is statistically independent, we recover the full Langevin equation (2.60)

$$\partial_t \rho(x,t) = \mathcal{D}\nabla^2\rho(x,t) + R[\rho(x,t)] + \eta(x,t), \tag{2.63}$$

where

$$R[\rho(x,t)] = -m^2\rho(x,t) - g\rho(x,t)^2. \tag{2.64}$$

As we discuss in Section 2.3.4, the choice $-m^2 = (\beta - \mu)$ and $g = \beta$, where β and μ are the transmission and recovery rates respectively, corresponds to the reaction-diffusion model of epidemic spreading with reaction given by the susceptible-infected-susceptible scheme. Under the stationary condition $\partial_t\rho(x,t) = 0$ and neglecting the spatial distribution of ρ, such that the diffusion term vanishes, the minimum of the functional (2.62) yields the inactive or absorbing stationary state $\rho(t=\infty) = 0$ and the active stationary state $\rho(t=\infty) = -m^2/g$, which is indeed the stable fixed point $\rho(\infty) = (\beta - \mu)/\beta$ of the susceptible-infected-susceptible model (see next Section). For epidemic spreading we will always consider the infected occupancy of each site as statistically independent and also, as for the upper-critical region of regular lattices where the noise is irrelevant, we will neglect the fluctuations introduced by the noise term $\eta(x,t)$ in the Langevin equation (2.60). As a first step we consider a spatially homogeneous order parameter $\rho(t)$ while in Section 2.3.3 and 2.3.4 we retain spatial inhomogeneity given by the dependence on the position x and consider the contact-network and the reaction-diffusion dynamics, respectively. As the critical point in statistical field theory is defined by the point of zero mass [212] $m^2 = 0$, the critical point of the reaction-diffusion model of epidemic spreading is defined by the point $\beta = \mu$ in the epidemiological parameters space. Other important models of non-equilibrium phase transitions that belong to different universality classes than that of directed percolation include the *Manna universality class* of *sandpile* [188] and *forest-fire* models [12, 64] displaying *self-organized criticality* [13] and *dynamical percolation*. The latter, also called *generalized epidemic process* [57, 58], is obtained as a generalization of directed percolation by including the effect of permanent immunization and corresponds to the susceptible-infected-removed scheme with diffusive coupling, see Section 2.3.4.

2.3.2. Mean field theory

Epidemic models generally assume that the population can be divided into different classes or compartments depending on the stage of the disease [7]. The three basic compartments are: healthy or susceptible (S), those who can contract the infection,

infected (I), those who contracted the infection and are contagious, and recovered or removed (R), those who are removed from the propagation process, either because they have recovered from the disease or because they have died. Additional compartments can be introduced to model other intermediate stages of the epidemic. Compartmental models can be extended to take into account vectors, such as mosquitoes, for diseases propagating through contact with an external carrier. Epidemic modeling describes the dynamical evolution of the contagion process within a population.

In order to understand the evolution of the number of infected individuals in the population as a function of time we have to define the basic individual-level processes that govern the transition of individuals from one compartment to another. Although epidemic spreading is best described as a stochastic reaction-diffusion process [262] governed by (2.63), the classic understanding of epidemic dynamics is based on the continuous-time limit of difference equations for the evolution of the average number of individuals in each compartment. This deterministic approach relies on a mean-field (homogeneous mixing) approximation, which assumes that the individuals in the population are well mixed and interact with each other completely at random, in such a way that each member in a compartment is treated similarly and indistinguishably from the others in that same compartment. Under this approximation, full information about the state of the epidemics is encoded in the total number X of individuals in the compartment $X \in \{S, I, R\}$ or, analogously, in the respective density $\rho^X = X/N$, where N is the population size. The time evolution of the epidemic is described by deterministic differential equations, which are constructed applying the law of mass action. Neglecting the recovery process, the average change in the infected population density is given by the product of the *force of infection*, i.e. the probability at which one susceptible individual may contract the infection in a single time step, times the susceptible population density. The distribution of the infectious period and the transmission probability can generally be estimated from clinical data. However, in a simplistic modeling scheme, the probability of transmission is often assumed to be constant. In this way, a discrete-time formulation defines a transmission and recovery probability per time step, the rates β and μ respectively.

Let us consider a constant population of N individuals with infinitely long-range interactions, i.e. we allow each individual to be able to interact with any other individual in the population at each time step of length dt. Within this scheme that neglects any geographical or demographic factor, the simplest compartmental model is the susceptible-infected (SI) model. The reaction in this case is

$$S + I \xrightarrow{\beta} I + I \tag{2.65}$$

where $S(t)$ and $I(t)$ are the number of susceptible and infected individuals respectively, satisfying the constraint $S(t) + I(t) = N$ at each time t, while βdt is the probability for a susceptible individual to become infected in the time interval dt. The force of infection

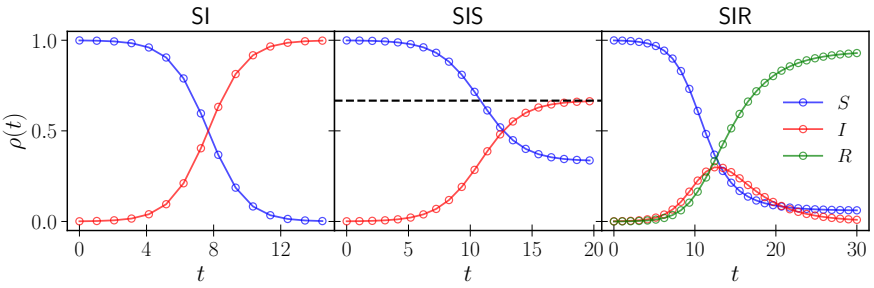


Figure 2.9: Epidemic curves for the SI (left) SIS (center) and SIR dynamics (right) in a population of $N = 1000$ individuals starting with a single infected $I(0) = 1$. Transmission and recovery rates are respectively $\beta = 0.9$ and $\mu = 0.3$ per time step. The dashed black line in the middle panel marks the stationary state $\rho^I(\infty) = (\beta - \mu)/\beta \approx 0.66$ that correspond to the stable fixed point of the SIS model. In all three cases the early stage of the dynamics is dominated by an exponential increase of the infection and the dynamics can be considered essentially linear. After the characteristic time $\tau = (\beta - \mu)^{-1}$ the non-linear effects are non-negligible and the curves rapidly saturate over the stationary state.

in the continuous-time limit is the product of the transmission rate β with the effective number of contacts per time step (in this case 1) times the fraction of infectious contacts $I(t)/N$, resulting in $\lambda(t) = \beta I(t)/N$. The number of infected individuals grows in time as $I(t + dt) = I(t) + \lambda(t)dtS(t)$ so the associated differential equation reads

$$\dot{I}(t) = \beta \frac{S(t)}{N} I(t). \tag{2.66}$$

The last equation can be interpreted as follows. The increment in the infected population $\Delta I(t) = I(t + dt) - I(t)$ equals the number of infected $I(t)$ times the fraction of susceptible becoming infected in the time dt, which is obtained multiplying the probability of becoming infected βdt times the fraction of susceptible $S(t)/N$. In terms of the infected density $\rho^I = I/N$ the dynamics becomes

$$\dot{\rho}^I = \beta(1 - \rho^I)\rho^I, \tag{2.67}$$

which describes a simple unconditioned *logistic growth* [251] of the infected population. The *epidemic prevalence* ρ^I is shown in the left panel of Figure 2.9. The stationary state is obtained imposing $\dot{\rho}^I = 0$, and it yields[12] a fully infected population $\rho^I = 1$ that corresponds to unitary carrying capacity of the logistic equation for population growth.

A straightforward generalization of the SI model is obtained by allowing temporary recovery from the disease, leading to the susceptible-infected-susceptible (SIS) model.

[12] This is the stable fixed point solution fo the active phase, while there is also the trivial unstable fixed point $\rho^I = 0$ of the absorbing phase, see previous Section.

The reaction in this case is

$$\begin{aligned} S + I &\xrightarrow{\beta} I + I \\ I &\xrightarrow{\mu} S \end{aligned} \tag{2.68}$$

The additional transition with respect to the SI model, $I \to S$, occurs when an infected individual recovers from the disease and returns to the pool of susceptible individuals at rate μ. The SIS model assumes that the disease does not confer immunity and individuals can be infected over and over again, undergoing the cycle $S \to I \to S$, which, under some conditions, can be sustained forever. The dynamics of the SIS model reads

$$\dot{\rho}^I = \beta(1 - \rho^I)\rho^I - \mu\rho^I, \tag{2.69}$$

with the normalization condition $\rho^S(t)+\rho^I(t) = 1$, that implies $\dot{\rho}^S = -\dot{\rho}^I$. The epidemic prevalence ρ^I for the SIS model is shown in the center panel of Figure 2.9. In this case the stationary state is an *endemic state*, characterized by a constant fraction of infected population $\rho^I(\infty) = (\beta - \mu)/\beta$ (dashed black line in Figure 2.9) which reduces to the stable equilibrium of the SI model for $\mu = 0$. By allowing for permanent immunization leads to the susceptible-infected-removed (SIR) model

$$\begin{aligned} S + I &\xrightarrow{\beta} I + I \\ I &\xrightarrow{\mu} R \end{aligned} \tag{2.70}$$

For any value of the transmission rate β and recovery rate μ, the SIR process will always asymptotically die after affecting a given fraction of the population which depends on the rates and on the initial condition $I(0)$. The dynamics become two-dimensional and reads

$$\begin{cases} \dot{\rho}^S = -\beta\rho^S\rho^I \\ \\ \dot{\rho}^I = \beta\rho^S\rho^I - \mu\rho^I \end{cases} \tag{2.71}$$

The evolution for the recovered compartment $\dot{\rho}^R = \mu\rho^I$ is decoupled and is obtained from the constraint $\rho^S(t) + \rho^I(t) + \rho^R(t) = 1$. The epidemic curves for the SIR model are shown in the right panel of Figure 2.9. In this case, the recovered density $\rho^R(t)$ behaves as the prevalence $\rho^I(t)$ of the SIS model, increasing logistically up to the stationary state value $\rho^R(\infty)$.

In both SIS and SIR model we have introduced the time scale μ^{-1} governing the self-recovery of individuals. We can think of two limiting cases: if μ^{-1} is smaller than the spreading time scale β^{-1}, then the process is dominated by recovery and by the decay into a healthy state. Instead when $\beta > \mu$, the spreading time scale is smaller than the

recovery time scale and the early dynamics reduces to the SI model. At the early stage of the epidemic we can assume $\rho^I \approx 0$ and we can linearize the equations to obtain for both the SIS and SIR model

$$\dot{\rho}^I \approx (\beta - \mu)\rho^I \qquad \Rightarrow \qquad \rho^I(t) = \rho^I(0)e^{t/\tau}. \tag{2.72}$$

The time scale of disease persistence $\tau = (\beta - \mu)^{-1}$ determines how fast the infected population grows in terms of the transmission and recovery rates. Contrary to the SI model, where one has $\mu = 0$ and thus $\tau > 0$ always, the characteristic time scale can become negative if $\mu > \beta$. When this happens the epidemic will not spread and will fade away on the time scale $|\tau|$. In other words the number of infected individuals grows exponentially if $\beta > \mu$ and decreases otherwise. This defines the *epidemic threshold*[13]

$$\mathcal{R}_0 = \frac{\beta}{\mu} > \widetilde{\beta}_c = 1, \tag{2.73}$$

where $\mathcal{R}_0$ is the *basic reproductive number*, defined as the average number of secondary infections caused by a primary case introduced in a fully susceptible population [7]. The rescaled transmission rate $\widetilde{\beta} = \beta/\mu$ incorporates the recovery in a compact control parameter, with a critical value in the homogeneous mixing approximation given by the epidemic threshold $\widetilde{\beta}_c = 1$. If $\mathcal{R}_0 < 1$, i.e. if a single infected individual generates less than one secondary infection, the relative size of the epidemics is negligibly small, vanishing in the thermodynamic limit of an infinite population. This concept is very general and the analysis of different epidemic models reveal in general the presence of a threshold behavior [139].

Spatial effects can be introduced in the above description by adding diffusive continuous terms or by considering patch models, as we will discuss in Section 2.3.4. Although a correct analysis of epidemic models should consider explicitly its stochastic nature, this is only important when dealing with small populations when the early stage of the process is dominated by the induced fluctuations. Finally, the classic deterministic approach assumes random and homogeneous mixing, where each member in a compartment is treated similarly and indistinguishably from the others in that same compartment and each individual is assumed to interact with a single randomly chosen individual per time step. In reality, however, each individual has his/her own social contact network over which diseases spread, usually differing from that of other members in a group or compartment. As we will see next, in a degree-block approximation we can compute the modified epidemic threshold, leading to a new definition of the basic reproductive number that depends explicitly on the topology of the underlying contact network.

[13] While in the deterministic scenario considered here the threshold condition $\mathcal{R}_0 > 1$ is both necessary and sufficient to have an epidemic outbreak, in real systems this is only necessary because of large fluctuations, that can be modeled introducing stochastic terms, at the early stage of the process.

2.3.3. Contact networks

Realistic models of epidemic spreading need to take into account the interaction patterns between individuals that can be conveniently modeled with a network. In contact networks we place single individuals at each node and let the classic compartmental model run on top of this determined interaction pattern. To each node is assigned the variable $\sigma_i(t) \in \{S, I, R\}$ and depending on the neighboring interaction of site i, its state might undergo the transition to a different state $\sigma_i'(t)$. As a first approximation we can assume that each node is in contact with the same number of nodes $\langle k \rangle$. This homogeneous assumption on contact networks, that becomes reasonable for networks with degree distributions $\mathcal{P}(k)$ converging to the Poissonian profile (2.11) and is exact for regular graphs, yields a simple rescaling of the transmission rate $\beta \to \beta \langle k \rangle$. Then the basic reproductive number is also rescaled as $\mathcal{R}_0 \to \langle k \rangle^{-1} \mathcal{R}_0$ so that the epidemic threshold condition (2.73) in this approximation turns into $\beta/\mu > \widetilde{\beta}_c = \langle k \rangle^{-1}$. This simple rescaling however, does not take into account that many networks exhibit very heterogeneous topologies. Generally, it is possible to show that $\mathcal{R}_0$ gets *renormalized* by fluctuations in the transmissibility or contact patterns as

$$\mathcal{R}_0 \to \mathcal{R}_0' = \phi\left(\mathcal{R}_0\right), \tag{2.74}$$

where ϕ is a positive and increasing function of the connectivity variance [214]. As we will see next, for the SIR model a good approximation is $\phi(\mathcal{R}_0) = \mathcal{R}_0 \left(\langle k^2 \rangle / \langle k \rangle - 1\right)$. This means that for heterogeneous networks we expect the fluctuations, and not the average degree, to play the main role in determining the epidemic properties.

In general it is necessary to go beyond the homogeneous-contact assumption. This is particularly important when the degree distribution follows a power law of the form (2.15), and topological fluctuations are present on virtually all scales. We can do this conveniently by grouping together the dynamical quantities with same degree k [23] as ρ_k^X for $X \in \{S, I, R\}$. Then the global averages are simply given by $\rho^X = \sum_k P(k) \rho_k^X$. Within this *degree-block approximation*, i.e. neglecting correlations between nodes' degrees, all nodes with same degree are statistically equivalent. The evolution equation for the infected density of the SIR model in this approximation read

$$\dot{\rho}_k^I = \beta k \Theta_k \rho_k^S \rho_k^I - \mu \rho_k^I. \tag{2.75}$$

The quantity $\Theta_k = \sum_{k'} P(k'|k) \rho_{k'}^I (k' - 1)/k'$ is the density of infected neighbors[14] of nodes with degree equal to k, which is defined in terms of the conditional probability for

[14] This expression for Θ_k valid for the SIR model, takes into account that a node cannot propagate the disease to the particular neighbor who originally infected it because the latter is necessarily not susceptible. Instead for the SIS model all neighbors of an infected node can receive the disease from it, including the one that originally passed the disease in the first place as we must consider also the possibility that it became susceptible again.

an edge to connect nodes with different degrees $P(k'|k)$. As the degrees are completely uncorrelated in this approximation the conditional probability factorizes as $P(k'|k) = k'P(k')/\langle k\rangle$ [23], hence $\Theta_k = \Theta$ is independent on the degree block k. By redefining Θ to incorporate also recovery as

$$\Theta = \sum_{k'} M_{kk'}\rho^I_{k'}, \qquad M_{kk'} = P(k')\frac{(k'-1)}{\langle k\rangle} - \frac{\delta_{kk'}}{k}\frac{\mu}{\beta}, \tag{2.76}$$

the linearization of the dynamics for the infected compartment yields $\dot{\rho}^I_k \approx \beta k\Theta$. The evolution equation $\dot{\Theta} = \sum_{k'} M_{kk'}\beta k'\Theta$ can easily be solved by separation of variables and gives

$$\Theta(t) = \Theta(0)e^{t/\tau}, \qquad \tau = \frac{\langle k\rangle}{\beta\langle k^2\rangle - (\beta+\mu)\langle k\rangle}. \tag{2.77}$$

Then for $\tau > 0$ the infected density $\rho^I(t)$ at the early stage increases exponentially fast. In networks with very heterogeneous connectivity the second moment $\langle k^2\rangle$ is very large and thus the outbreak time scale τ is very small signaling a very fast spreading. In particular, as we showed in Section 2.1.3, scale-free networks with $P(k) \sim k^{-\gamma}$ and $2 < \gamma \leq 3$ have $\langle k^2\rangle \to \infty$ in the thermodynamic limit while $\langle k\rangle$ stays finite. Therefore in uncorrelated scale-free networks we face a virtually instantaneous rise of the epidemic with $\tau \to 0$. This fact can be physically explained by the presence in scale-free network of strong hubs that can spread very rapidly the disease following a cascade process of decreasing degree classes [23]. In order to ensure an epidemic outbreak the condition $\tau > 0$ leads to the striking result for the epidemic threshold[15]

$$\frac{\beta}{\mu} > \widetilde{\beta}_c = \frac{\langle k\rangle}{\langle k^2\rangle - \langle k\rangle}, \tag{2.78}$$

where as before $\widetilde{\beta} = \beta/\mu$ is the control parameter of the epidemic phase transition. This result is precisely the threshold condition (2.73) for the *renormalized* reproductive number

$$\mathcal{R}_0 = \left(\frac{\langle k^2\rangle}{\langle k\rangle} - 1\right)\frac{\beta}{\mu}, \tag{2.79}$$

that takes into account the networked interaction between the system components.

Contrary to the mean-field scenario, the critical point (2.78) for contact networks implies that very heterogeneous networks, such as scale-free networks where $\langle k^2\rangle$ diverges with the system size, there is a null epidemic threshold. Importantly, this is a general result that also holds for real finite networks where the finite size effects introduced by

[15] For the SIS model an analogous derivation yields $\widetilde{\beta}_c = \langle k\rangle/\langle k^2\rangle$.

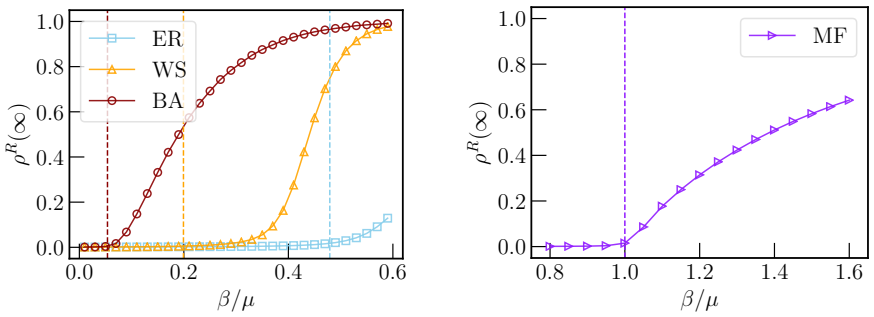


Figure 2.10: Final outbreak size $\rho^R(\infty)$ as a function of the control parameter β/μ for SIR contact-network averaged over 10^2 realizations and over all source seeds for artificial networks (left panel) each consisting of $N = 1000$ nodes: ER (light-blue) with edge-creation probability $p = 0.002$, WS (orange) with edge-rewiring probability $p = 0.02$ and $2m = \langle k \rangle = 6$ neighbors per node and BA (dark-red) with $m = 5$ new edges per time step. In violet (right panel) the curve for the mean field (MF) model with the homogeneous mixing assumption. The vertical dashed lines mark the corresponding epidemic thresholds. For the the contact networks the degree-block approximation (2.78) yields $\widetilde{\beta}_c^{\mathsf{ER}} \approx 0.48$, $\widetilde{\beta}_c^{\mathsf{WS}} \approx 0.20$ and $\widetilde{\beta}_c^{\mathsf{BA}} \approx 0.05$, respectively, while (2.73) defines the MF threshold $\widetilde{\beta}_c^{\mathsf{MF}} = 1$ (dashed violet).

the cutoff induce an epidemic threshold that eventually approaches zero at increasing sizes [216]. This remarkable result obtained in a degree-block approximation is well confirmed from empirical observation of real networks and has profound implications for disease eradication making scale-free networks the ideal environment for the spreading of diseases. In the case of uniform immunization it is easy to show that the introduction of a density of immune individuals ρ^R is equivalent to rescale $\mathcal{R}_0$ by the probability that any node is not immune $(1-\rho^R)$. The critical value ρ_c^R that corresponds to the epidemic threshold of the model (2.78) must then satisfy $(1-\rho_c^R)\beta/\mu = \langle k \rangle / (\langle k^2 \rangle - \langle k \rangle)$. Then as the fluctuations increase only the virtually complete immunization of the network, i.e. $\rho_c^R = 1$, ensures an infection-free stationary state.

We conclude the Section by highlighting the correspondence between SIR contact networks and percolation theory. The connection between the static properties of the SIR model and (isotropic) bond percolation on the lattice was recognized long ago [117, 130, 271, 241]. In the context of epidemics on complex networks, the mapping has been studied in detail by Newman [198]. Considering a SIR model with uniform infection time τ, i.e., where infected nodes become recovered at time $\tau = \mu^{-1}$ after contracting the infection, and infection rate β, the transmissibility p is defined as the probability that the infection will be transmitted from an infected node to a connected susceptible neighbor before recovery takes place. In the continuous-time limit the transmissibility

can be computed as

$$p = 1 - \lim_{\delta t \to 0} (1 - \beta \delta t)^{\tau/\delta t} = 1 - e^{-\beta/\mu}. \tag{2.80}$$

The set of recovered nodes generated by an SIR epidemic outbreak originated from a single node is nothing else than the cluster of the (isotropic) bond percolation problem with occupation probability p, to which the initial node belongs. The correspondence in this case is exact: all late-time static properties of the SIR model can be derived as direct translations of the geometric properties of the percolation problem. The previous argument can also be extended to a non-uniform infection time. More realistically, we assume that infection times τ_i and transmission rates β_{ij} vary between individuals [271, 241]. This implies that the transmissibility p_{ij} depends on the specific edge. One possible approach, that reduces to the solution of the homogeneous case [198], is to neglect fluctuations, and replace p_{ij} by its mean value $\langle p_{ij} \rangle$, where the average is taken over the infection time and transmission rate distributions $\mathcal{P}(\tau)$ and $\mathcal{P}(\beta)$.

2.3.4. Metapopulations

So far we considered spreading processes without a definite interaction pattern or where each node of the interaction network corresponds to a single individual of the population. A different approach consists in considering nodes as patches where multiple individuals can be located and then eventually move along the links connecting the nodes. Examples of such systems are provided by mechanistic *reaction-diffusion* models [262, 191] (see Section 2.3.1) where particles represent people moving between different locations or by the routing of information packets in technological networks. This framework is particularity useful to model the spreading of epidemic in spatially-structured *subpopulations* [233], such as city locations, urban areas, or geographical regions and defines the *metapopulation model*. Individuals are divided into classes denoting their state with respect to the modelled disease, as for homogeneous mixing compartmental models, and the reaction processes account for the possibility that individuals in the same location may get in contact and change their state according to the infection dynamics.

Mathematically, we describe the epidemic process in a metapopulation by considering N subpopulations (nodes) connected by E weighted edges. The structure of the metapopulation is then defined by the weighted adjacency matrix W_{ij} (as given from real data) that gives the number of people traveling from subpopulation i to subpopulation j in a time step. The strength $s_i = \sum_k W_{ik}$ gives the total out-flux of subpopulation i. In general the subpopulation size N_j is an independent variable and the total number of individuals in the metapopulation is $\mathcal{N} = \sum_j N_j$. The traveling flux between subpopulations can be express as $W_{ij} = Q_{ij} N_i$, where Q_{ij} are the traveling rates in a continuous-time description defined by (2.37) which represent the conditional probability per time step of a randomly chosen individual to jump from location i to location j.

In the absence of real data a common approach is to assume a gravity-like interaction between subpopulations [266, 101] so that $W_{ij} \sim N_i N_j / D_{ij}^{\alpha}$, where α is a free parameter to be tuned and D_{ij} is the geographical distance between subpopulations i and j.

For the SIR reaction (2.71) we denote by S_j, I_j and R_j the number of individuals who belong to subpopulation j who are in the susceptible, infected, and removed state, respectively. The subpopulation size is $N_j(t) = S_j(t) + I_j(t) + R_j(t)$. The correspondent normalized quantities are denoted as $\rho_j^X = X_j/N_j$, where $X \in \{S, I, R\}$. Then, the quantity $\rho_j^I(t)$ can be viewed as the probability that node j is infected at time t. The movement of a host between subpopulations is governed by the reaction kinetics $X_k \xrightarrow{Q_{ki}} X_i$, i.e. $\dot{X}_i = \sum_k X_k Q_{ki}$. By splitting the ith term in the sum and using the definition of exit rate (2.36), we obtain

$$\dot{X}_i = \sum_{k \neq i} X_k Q_{ki} - \sum_{k \neq i} X_i Q_{ik}, \tag{2.81}$$

i.e. the master equation (2.39) for the number of individuals in compartment X. We always assume that any initial condition for N_i satisfy the stationary state and since (2.81) for $X_i = N_i$ yields

$$\dot{N}_i = \sum_k \left(W_{ki} - W_{ik} \right), \tag{2.82}$$

the stationary condition is equivalent to symmetric travels between subpopulations. The symmetry of the weighted adjacency matrix is always satisfied in large-scale real transportation networks to a very high degree of accuracy [22]. For symmetric unweighted networks for which (2.31) holds, this also implies that the system satisfies detailed balance

$$W_{ij} = Q_{ij} N_i = Q_{ji} N_j = W_{ji}, \tag{2.83}$$

where now $N_i / \sum_l N_l$ is the stationary distribution over the set of nodes. Note that the previous relation holds only for $i \neq j$, since the diagonal rates $Q_{ii} = -\sum_{k \neq i} Q_{ik}$ are negative by construction while we always assume that $W_{ii} = 0$, i.e. that there are no self-edges. By inserting (2.83) into the master equation (2.81), we get a generalization to the the diffusion equation on graphs[16] (2.41)

$$\dot{\rho}_i^X = \sum_k \left(\rho_k^X \frac{W_{ki}}{N_i} - \rho_i^X \frac{W_{ik}}{N_i} \right) = \sum_k \frac{W_{ik}}{N_i} \left(\rho_k^X - \rho_i^X \right) = -\sum_k \frac{L_{ik}}{N_i} \rho_k^X, \tag{2.84}$$

[16] The ith term in the sum can be re-included when going from rates to weights since we are assuming that no self-edges are allowed i.e. $W_{ii} = 0$ for all i.

where

$$L_{ij} = \delta_{ij} s_i - W_{ij} \tag{2.85}$$

is the *weighted Laplacian* that generalizes (2.42) with $s_i = \sum_k W_{ik}$.

In the following we show first that there is an equivalent formulation of the above result that does not require the knowledge of the subpopulation size. Secondly, we combine the diffusion equation with a reaction term $R[\rho^X(x,t)]$ from the mean-field theory of compartmental models to get the full reaction-diffusion scheme of the metapopulation model. To remove the dependence on N_i we assume that the subpopulation size and the strength are proportional via a node-independent constant, the *diffusion rate* α. The local diffusion rate is defined as the exit rate (2.36). Then, if the exit rate

$$q_i = \sum_{k \neq i} Q_{ik} = \sum_k \frac{W_{ik}}{N_i} = \frac{s_i}{N_i}, \tag{2.86}$$

is independent on the particular node i, we can define the *diffusion rate* as

$$\alpha = q_i, \ \forall i. \tag{2.87}$$

The latter is also equal to the fraction of the total metapopulation traveling per unit time $\alpha = \sum_i s_i / \sum_i N_i = \sum_{ij} W_{ij} / \sum_i N_i$. Using that $W_{ki} = W_{ik}$, multiplying and dividing by the exit rate $q_i = \alpha$ in (2.84) we can cast the diffusion in the form $\dot{\rho}_i^X = \alpha \sum_k P_{ik} \left(\rho_k^X - \rho_i^X \right)$, where

$$P_{ij} = \frac{W_{ij}}{\sum_l W_{il}} = \frac{Q_{ij}}{\sum_{l \neq i} Q_{il}} \tag{2.88}$$

is the transition matrix. The full metapopulation model is obtained by combining diffusion with the compartmental reactions (2.69) or (2.71) in a unified picture that reads

$$\begin{cases} \dot{\rho}_i^S = \alpha \sum_k P_{ik} \left(\rho_k^S - \rho_i^S \right) - \beta \rho_i^S \rho_i^I + \chi \rho_i^I \\ \\ \dot{\rho}_i^I = \alpha \sum_k P_{ik} \left(\rho_k^I - \rho_i^I \right) + \beta \rho_i^S \rho_i^I - \mu \rho_i^I \end{cases} \tag{2.89}$$

where $\chi = \mu$ for SIS and $\chi = 0$ for SIR. When the approximation of uniform exit rates $\alpha = q_i = \sum_{l \neq i} Q_{il}$ is not adequate, we simply replace in the previous system of equations, αP_{ik} with the rates Q_{ik}, which are obtained from the data or by normalizing the travel fluxes W_{ik} by the source subpopulation size N_i.

It is important to stress that the metapopulation model (2.89) considers a simplified mechanistic approach with a Markovian assumption in which individuals are not labeled

according to their original subpopulation, and where at each time step the same traveling probability applies to all individuals in each subpopulation, without any memory of their previous locations. Instead at each time step the movement of individuals is given according to the matrix Q_{ij} that expresses the rate at which any individual in the subpopulation i is traveling to the subpopulation j. Furthermore the model considered here is fully deterministic, since it considers only expectation values, while both epidemic spreading and travel of individuals are inherently stochastic processes. The number of infected individuals is treated as a continuous variable and, even if the initial condition at $t = 0$ consists of one single infected individual in node i_0, all other nodes have a non-zero density of infected at any time $t > 0$. Given its limitations, this approach is nevertheless extensively used for large populations where the traffic W_{ij} between subpopulations is known and the subpopulations are large enough to safely assume homogeneous mixing without additional stochastic effects.

3

Reaction-Diffusion on Random Networks

"*Science is the belief in the ignorance of experts.*"

–Richard Feynman

Contents

The metapopulation model described in Section 2.3.4, has been successfully used to simulate global epidemic outbreaks in spatially embedded populations, such as cities and urban areas, interacting with each other [74, 75, 261]. In this Chapter, we want to characterize the most general mobility profile and model real human motion using *random networks*. Although being complicated objects, random networks can often be characterized by a small number of parameters due to the universal behavior of a whole ensemble. We therefore demonstrate that knowledge of the universality class of a transport network, described by some scaling exponent, is enough to extract crucial information about the infection process. In addition, to properly understand transport

in real systems it is important to embed the network into the geographic space, i.e. one has to consider spatial networks [25].

Often, the analytical treatment of an ensemble of random entities is necessary and more viable than a single representation. Relevant examples include amorphous materials such as spin glasses [194] and protein secondary structures [211, 137]. Several methods to deal with disorder in equilibrium systems are known, the most famous of which is probably the replica method [249] and the theoretical framework to treat disordered systems at equilibrium is a reach and established field of research [86]. For such systems, one is interested in computing the *quenched average* of macroscopic thermodynamic quantities such as the free energy (i.e. of the logarithm of the partition function) over the microscopic disorder realizations (e.g. of some couplings between the degrees of freedom). As opposed to the annealed average, which is appropriate when the disorder is not fixed and is allowed to fluctuate with the dynamical degrees of freedom, the quenched average is much harder to compute. For a thermodynamic system, the annealed average is taken directly over the partition function, which is itself a sum over system configurations of the degrees of freedom (e.g. the spin in a magnet), while the quenched average is taken on the logarithm of the partition function. Thus, the quenched average considers the disorder of the system as "frozen" in time.

In the simplified case considered here, the quenched average has to be performed in a "geometrical" rather than thermodynamical system. In particular, we leverage *effective medium theory* (EMT) to replace an ensemble of random networks with a single deterministic representative network that retains, on average, the same macroscopic properties of the ensemble in a similar fashion to the mean-field approximation in statistical mechanics [212]. This deterministic network, called the effective medium, is characterized by an effective diffusion coefficient that describes transport in the ensemble. EMT is not a "blind" average of the transition rates, rather it is determined in a self-consistent manner. Would a link in the effective medium be replaced by its random original, the transport flux along this particular link would not change on average. Hence EMT is particularly suited for systems with independent links.

In this Chapter, we employ EMT to characterize epidemic spreading in structured random metapopulations. The spatial embedding of the networks is a crucial requirement to model human mobility, as two topologically adjacent nodes (e.g. airports) may be geographically very far apart.

Empirical observations have pointed out that human motion lacks a definite scale [48, 129] and features *long-range* connections which have been a major limitation for EMT. Recently [256], EMT has been extended to overcomes this restriction and provides an analytical technique to deal with random spatial networks with scale-free transition rates. In the following, we demonstrate that and how EMT can be used to extract relevant epidemiological quantities, such as the infection prevalence, in random spatial networks with long-range connections. Contemporary fields where our proposed theory may become relevant are epidemic spreading in the global mobility network [236, 72, 46],

or dispersal phenomena in the biological contexts [132].

The Chapter is organized as follows. In Section 3.1 we review effective medium theory in the broader context of random resistor networks. A description of ballistic spreading in deterministic metapopulations is outlined in Section 3.2. We leverage the Feynman-Kac equation to derive a bound for the diameter of the infected cluster of nodes in the deterministic geographic setting. Then, in Section 3.3 we proceed to show that this estimate is as well realized in random models, where it can be computed from the effective medium approximation. The effective medium for the particular case considered here, i.e. when the transition rates follow a power law is derived analytically. Finally, we confirm with numerical experiments our theory by direct simulations of global outbreaks using a metapopulation model with long-range connectivity in the geographic setting of real human motion. The results presented in this Chapter are discussed in [150].

3.1. Effective medium theory

EMT is an ancient tool, originally developed to determine material properties in continuous and lattice model of heterogeneous media [51, 66, 158]. EMT is routinely applied in systems with short-range connections, but only recently was generalized to systems with long-range connections [256]. The latter theory is employed here to a disordered reaction-diffusion system with long-range connections decaying as power laws with the geographical distance. This setup serves as a proxy for epidemic spreading for a general mobility network. In the following we review EMT for the a network of random resistors, extending the derivation of Kirkpatrick [158] from regular lattices to a general network topology as done in [256].

Let us consider a network where at each edge is placed a random conductance, defined as the inverse resistance $G_{ij} = 1/R_{ij}$, drawn independently from some probability distribution $\mathcal{P}(G_{ij})$. The relation between the conductance G_{ij} and the current I_{ij} flowing from i to j is given by Ohm's law

$$I_{ij} = G_{ij} V_{ij} = G_{ij}(v_j - v_i), \tag{3.1}$$

where v_i is the electric potential between node i and the ground. In terms of the associated electric charges $q_i(t) = C_i v_i(t)$, where C_i is the (time-independent) capacitance between node i and the ground, the previous equation summed over all adjacent nodes $\{j\}$ to i, gives $\dot{q}_i(t) = \sum_j I_{ij}$. Then, we obtain the master equation of the density (2.84) for the electric potential $v_i(t)$

$$\dot{v}_i(t) = \sum_j \frac{G_{ij}}{C_i}(v_j(t) - v_i(t)), \tag{3.2}$$

where the symmetric conductances G_{ij} correspond to the symmetric flux of individuals

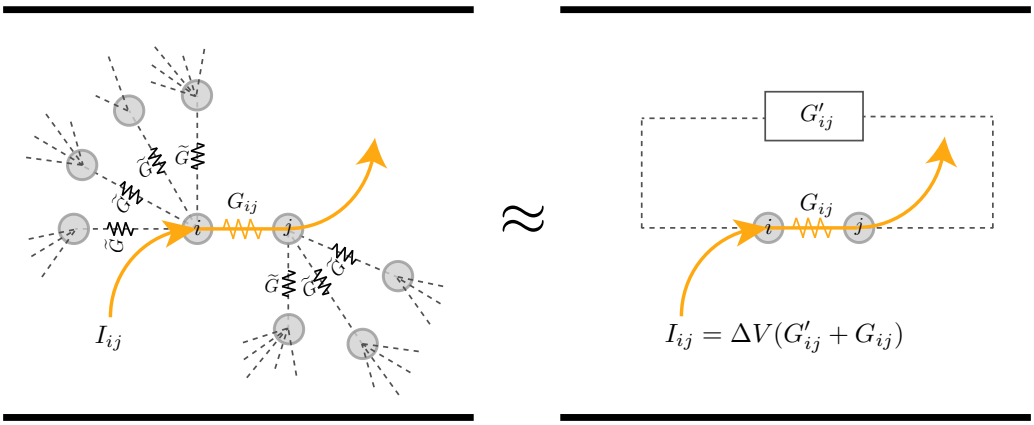


Figure 3.1: Effective medium for the random resistor network. In the homogeneous network a resistor $\widetilde{G}$ is replaced by a random value G_{ij} and a current I_{ij} is introduced at node i so that the potential difference V_{ij} between i and j is restored to the original homogeneous value $\widetilde{V}$ (left panel). The extra voltage $V_{ij} = \Delta V + \widetilde{V}$ introduced by the current I_{ij} is computed from the value of the parallel conductance G'_{ij} of the network between points i and j when G_{ij} is absent (right panel). Note that the graph in the left panel is a tree only for visualization purposes, and in general there are edges connecting the neighbors of i with j and with its neighbors.

W_{ij} traveling between i and j, C_i to the (constant) number of individuals N_i at site i and G_{ij}/C_i is the associated transition rate $Q_{ij} = W_{ij}/N_i$.

Let us now consider an homogeneous medium, for which we have $G_{ij} = \widetilde{G}$ for all edges (i, j) and the respective voltages $V_{ij} = \widetilde{V}$ are also constant. If we replace one of the homogeneous conductances, say at the edge (i, j), with the conductance G_{ij}, then the voltage difference amounts to $\Delta V = V_{ij} - \widetilde{V}$. The effective medium is defined as the value $\widetilde{V}$ such that $\overline{\Delta V} = 0$, where $\overline{(\cdots)}$ is the (quenched) average over the disorder with distribution $\mathcal{P}(G_{ij})$. If we now connect an additional current source (a battery) to the nodes i and j, we can tune the external current I_{ij} until the potential difference V_{ij} between i and j gets equal to the original homogeneous value $\widetilde{V}$, see Figure 3.1. After introducing the battery, the potentials of all nodes are restored and all currents through other resistors are equal to $\widetilde{I} = \widetilde{G}\widetilde{V}$, as in the case when $G_{ij} = \widetilde{G}$. This means that the whole additional current flows through G_{ij} and does not redistribute over other resistors. Therefore the total current I_{ij} flowing from i to j (including the background homogenous current $\widetilde{I}$) is related to G_{ij} via

$$I_{ij} = (G_{ij} - \widetilde{G})\widetilde{V}. \tag{3.3}$$

If we switch off this current, i.e. add a current of strength I_{ij} flowing in the opposite

direction (with a minus sign), we can write down an additional equation in terms of the parallel circuit to the random conductance G_{ij}. Indeed, since the equivalent circuit consist of the two conductivities, G_{ij} of the resistor considered and G'_{ij} of the rest of the system, switched in parallel (right panel in Figure 3.1), the voltage net change $\Delta V = V_{ij} - \widetilde{V}$ is equivalent to

$$\Delta V = \frac{I_{ij}}{G'_{ij} + G_{ij}} = \widetilde{V} \frac{G_{ij} - \widetilde{G}}{G'_{ij} + G_{ij}}, \tag{3.4}$$

where we have used (3.3). The self-consistency requirement of EMT then suggests that ΔV has to vanish, on the average, if $\widetilde{G}$ is chosen correctly for any value of the homogeneous voltage $\widetilde{V}$, i.e.

$$0 = \overline{\frac{\widetilde{G} - G_{ij}}{G'_{ij} + G_{ij}}} = \int dG_{ij}\ \mathcal{P}(G_{ij}) \frac{\widetilde{G} - G_{ij}}{G'_{ij} + G_{ij}}. \tag{3.5}$$

The remaining task is the calculation of the parallel-circuit conductance G'_{ij}. In the case of a regular lattice in d dimensions, a simple calculation yields [158] $G'_{ij} = (d-1)\widetilde{G}$ and the self-consistent condition (3.5) can easily be solved for $\widetilde{G}$, for a given disorder distribution $\mathcal{P}(G_{ij})$.

As a next step, let us consider the general scenario with no assumption on the network topology and with effective medium conductances $\widetilde{G}_{ij}$ that depend on the particular edge (i, j). We only assume that both the ensemble of random conductances and the effective medium conductances are symmetric, so that (3.2) holds. The self-consistency equations (3.5) can be rewritten in the general case by evaluating G'_{ij} in terms of the *total conductance* of the pure effective medium $G^*_{ij} = G'_{ij} + \widetilde{G}_{ij}$ (obtained as for (3.4) from switching in parallel $\widetilde{G}_{ij}$ with the rest of the medium) as [256]

$$0 = \overline{\frac{\widetilde{G}_{ij} - G_{ij}}{G^*_{ij} - \widetilde{G}_{ij} + G_{ij}}} = \overline{\frac{R^*_{ij}(\widetilde{G}_{ij} - G_{ij})}{1 - R^*_{ij}(\widetilde{G}_{ij} - G_{ij})}}, \tag{3.6}$$

where $R^*_{ij} = 1/G^*_{ij}$ is the total resistance of the pure effective medium. As shown in [256], by using the master equation (3.2) for a general effective medium $\{\widetilde{G}_{ij}\}$, the total resistance R^*_{ij} is a known quanityt in graph theory and is equivalent to the *graph resistance distance* [16] defined as

$$R^*_{ij} = \widetilde{L}^{-1}_{ij} + \widetilde{L}^{-1}_{ji} - \widetilde{L}^{-1}_{ii} - \widetilde{L}^{-1}_{jj}. \tag{3.7}$$

The latter is computed from the (pseudo-)inverse of the Laplacian matrix $\widetilde{L}_{ij}$ of the effective medium graph where conductances are replaced with travel fluxes.

The self-consistent condition (3.6), although it is not exact, has been known and used successfully for systems far away from the percolation threshold. Importantly, we made no assumptions on the particular disorder distribution nor on the topology of the network. The key ingredient to solve (3.6) for $\widetilde{G}_{ij}$ is in principle only the knowledge about the total resistance of the effective medium (3.7). A crucial requirement for EMT to work is that any link that is present in the random networks must also be present in the effective medium (albeit with possibly different weight) and with the same spatial scaling. Note that the specific form of the effective medium $\widetilde{G}_{ij}$ is mostly arbitrary, as long as the condition given above is respected.

In the thermodynamic limit we can further simplify the self-consistent condition (3.6) to obtain an asymptotic effective medium which has a simple form. Remembering that $R^*_{ij} = 1/G^*_{ij}$ is the total effective medium resistance between i and j, while $\widetilde{R}_{ij} = 1/\widetilde{G}_{ij}$ is the single resistor placed on the edge (i,j), we have $R^*_{ij}/\widetilde{R}_{ij} \leq 1$, where the equality is for a network of two nodes only. Indeed, in general the total resistance consists of $\widetilde{R}_{ij}$ and possibly many other parallel resistors so that $1/R^*_{ij} \approx \sum_{(i,j)} 1/\widetilde{R}_{ij} \gg 1/\widetilde{R}_{ij}$. The expansion of the geometric series in (3.6) for $R^*_{ij}/\widetilde{R}_{ij} \ll 1$ yields [256]

$$0 = \overline{\frac{R^*_{ij}/\widetilde{R}_{ij}\left(1 - G_{ij}/\widetilde{G}_{ij}\right)}{1 - R^*_{ij}/\widetilde{R}_{ij}\left(1 - G_{ij}/\widetilde{G}_{ij}\right)}} \approx \frac{R^*_{ij}}{\widetilde{R}_{ij}} \overline{\left(1 - \frac{G_{ij}}{\widetilde{G}_{ij}}\right)}. \tag{3.8}$$

Thus, in the limit of high connectivity, i.e. far from the percolation threshold, the effective medium solution is simply the disorder average of the single conductance

$$\widetilde{G}_{ij} = \overline{G}_{ij}, \quad \forall (i,j). \tag{3.9}$$

This general result of EMT, that is valid for highly connected networks, constitute the main ingredient for our analytical derivations that we outline in Section 3.3 in the context of transport in metapopulation networks, where the random conductances $\{G_{ij}\}$ are replaced by the fluxes $\{W_{ij}\}$.

3.2. Spreading in deterministic networks

3.2.1. Metapopulation model and Feynman-Kac estimate

We will demonstrate our idea for the contact process (directed percolation) described in Section 2.3, i.e. using the SIS reaction-diffusion model with transmission and recovery rates β and μ respectively. We assume that all subpopulations (located at the nodes of the metapopulation network) have the same amount of individuals[1] $N_x = \mathcal{N}$ for all

[1] From here we use the continuous-space notation for the nodes index $x, y, z, \ldots$ as opposed to the standard network notation $i, j, k, \ldots$ since it makes clearer the next derivations.

x and are placed at equal distances on the infinite line $\mathbb{Z}$. We allow individuals to travel not only from x to adjacent subpopulations located at $x \pm 1$, but also to distant subpopulations at $x \pm \xi$ with $\xi \in \mathbb{N}$, in a similar fashion to the WS model (see Section 2.1.3). The number of people traveling from x to y per unit time is the product of the rate Q_{xy} with the x subpopulation size $\mathcal{N}$. We also assume the rates to be symmetric $Q_{xy} = Q_{yx}$, so that the flux of travelers in one direction $W_{xy} = \mathcal{N}Q_{xy}$ is balanced by the flux in the opposite direction, see (2.83). The assumption of symmetric fluxes is well confirmed from empirical observation in real air-transportation networks [21]. We stress that in this setting we have a well defined geometry, given by $\mathbb{Z}$ and the node index x denotes a position in *Euclidean* space. Therefore, a link that stretches only a unit distance in the topological sense, may connect two very remote nodes in geographical space. These *long-range* connections are our model for the fast superdiffusive transport characteristic of human mobility [48, 129]. In Figure 3.2 we illustrate the interplay between reaction and diffusion in our model with the embedding of the one-dimensional array of subpopulations with long-range connections on a two-dimensional surface such as the geographical space of human mobility at the global scale.

We denote with Ω_{xy} the *transport operator*, which accounts for diffusion in the metapopulation, defined as the negative weighted Laplacian matrix (2.85) normalized by the subpopulation size

$$\Omega_{xy} = -L_{xy}/\mathcal{N} = Q_{xy} - \delta_{xy} \sum_z Q_{xz}. \tag{3.10}$$

Let $p_x(t)$ denote the fraction of the population on site x at time t. In analogy with the resistor network model, by identifying the nodes potential $v_x(t)$ with the distribution $p_x(t)$ and the edge conductances G_{xy} the travel fluxes W_{xy}, the time evolution of $p_x(t)$ together with the initial condition $p_x(t=0) = \delta_{x,0}$ is described by the master equation (3.2) and reads

$$\dot{p}_x(t) = \sum_y \Omega_{xy} p_y(t) = \sum_y Q_{xy}(p_y(t) - p_x(t)). \tag{3.11}$$

The symmetry in the travel fluxes assures that, not only the the total number of individuals in the metapopulation but also the subpopulation size $\mathcal{N}$ remains constant during the dynamics, see (2.82). We denote with $\rho_x(t)$ the fraction of infected subpopulation x at time t. The evolution equation for ρ_x is obtained by adding a reaction term $R[\rho_x(t)]$ to the diffusion equation (3.11), where for the contact process defined by (2.64) we have

$$R[\rho_x(t)] = \beta(1 - \rho_x(t))\rho_x(t) - \mu\rho_x(t). \tag{3.12}$$

The full dynamics reads

$$\dot{\rho}_x(t) = \sum_y \Omega_{xy}\rho_y(t) + R[\rho_x(t)], \tag{3.13}$$

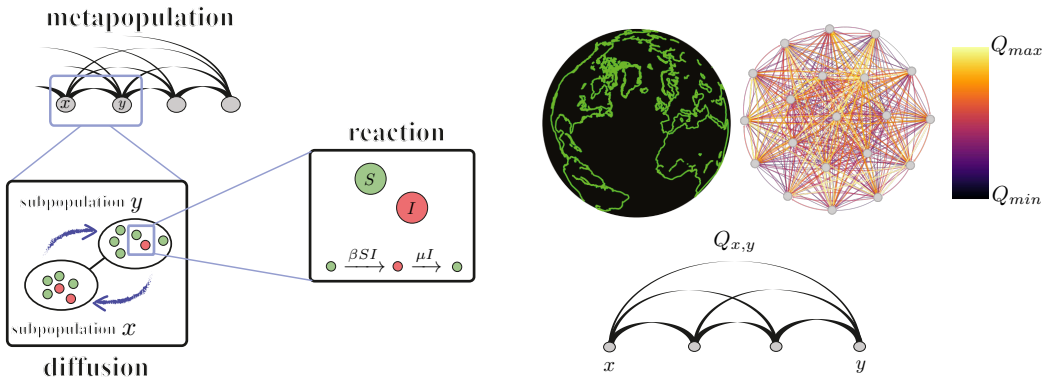


Figure 3.2: Left panel: scheme of the one-dimensional contact process. The dynamics is regulated by two different time scales, the one of diffusion, corresponding to the subpopulation layer, and the reaction, governed by the SIS infection dynamics at the individual layer. Right panel: illustration of a sample metapopulation network consisting of $N = 20$ subpopulations with symmetric transition rates Q_{xy}. The graph is constructed from a one-dimensional ring topology by adding all connections between nodes. This allows the embedding with a plane surface such as the geographical space of the global mobility network. The edge color and size scales accordingly with the values of each transition rate.

with initial condition $\rho_x(t=0) = c_0\delta_{x,0}$, given that a fraction c_0 is the initially infected in the seed subpopulation (node 0).

Our arguments in in what follows go parallel to those of [186]. Using the Feynman-Kac equation [110], we can write the formal solution $\rho_x(t)$ as an expectation value in a similar fashion to the path-integral formulation of quantum mechanics [242]

$$\rho_x(t) = \left\langle \rho_{x+X(t)}(0) \exp\left[\int_0^t dt' \frac{R[\rho_{x+X(t')}(t')]}{\rho_{x+X(t')}(t')}\right]\right\rangle . \tag{3.14}$$

Here, $X(t)$ is a random walk on $\mathbb{Z}$, that starts at $x = 0$ and evolves according to the master equation (3.11), and the average $\langle \ldots \rangle$ is taken over all such random-walk configurations. Hence, the transition probabilities for $X(t)$ are given by (2.37). Although the expectation value (3.14) is a complicated self-consistent equation that can hardly be solved for $\rho_x(t)$, it provides a very useful estimate. Indeed, since $\rho_x(t) \geq 0$, we have the obvious estimate

$$\frac{R[\rho_x(t)]}{\rho_x(t)} = \beta - \mu - \beta\rho_x(t) \leq \beta - \mu. \tag{3.15}$$

Plugging this inequality into (3.14) and identifying the transition probability $\langle \rho_{x+X(t)}(0)\rangle =$

$c_0 \langle \delta_{0,x+X(t)} \rangle = c_0 p_x(t)$, we obtain

$$\rho_x(t) \leq c_0 e^{(\beta-\mu)t} p_x(t). \tag{3.16}$$

This shows that $\rho_x(t)$ has an exponentially growing profile in time modulated by the *free diffusion*[2] of agents encoded in $p_x(t)$. Then, if we can find an analytical solution $p_x(t)$ for the free motion of agents, the inequality can be inverted for x to estimate the infection spreading front. To do so, we consider the level set, i.e. the set of nodes $\{x\}$, where the fraction of infected is higher than some threshold value c. The diameter of the set

$$\omega(t) \equiv \text{diam}\{x | \rho_x(t) \geq c\}, \tag{3.17}$$

determines the largest distance between any two nodes that belong to the level set. The quantity $\omega(t)$ is then the analogous of the epidemic prevalence, as measured with respect to an empirical precision given by the value c, on the ring with N nodes and metric

$$|x - y| = \min(|x - y|, N - |x - y|)\,. \tag{3.18}$$

3.2.2. Ballistic versus exponential spreading

As a first step, let us consider a simple random walk on the line (with equal probability to jump in the right or on the left) with short-range connections only. We assume that individuals travel from a subpopulation x to adjacent subpopulations $x \pm 1$ with a constant transition rate (probability per time step) $Q_{xy} = Q$, independently on the node location. Then, $Q\rho_x$ infected individuals will travel from x to $x + 1$ and to $x - 1$ per unit time. In this case, as for (2.41), the transport operator $\Omega_{x,y}$ that describes diffusion of simple random walks is equal to the negative (unweighted) Laplacian on the line multiplied by the diffusion coefficient [262]

$$\Omega_{x,y} = -\mathcal{D}\left(2\delta_{x,y} - \delta_{x,y-1} - \delta_{x,y+1}\right), \tag{3.19}$$

where $\mathcal{D} = Q$. If one assumes that the free diffusion is Gaussian, with $p_x(t)$ given by the continuous-time version of (2.26) and diffusion coefficient Q, plugging this in (3.16) yields

$$c \leq \rho_x(t) \leq c_0 \frac{e^{(\beta-\mu)t}}{\sqrt{4\pi Q t}} \exp\left(-\frac{x^2}{4Qt}\right), \tag{3.20}$$

which, after partially solving for x, gives

$$\frac{x^2}{4Qt} \leq (\beta - \mu)t - \ln\left(\frac{c}{c_0}\sqrt{4\pi Q t}\right). \tag{3.21}$$

[2]Note that here we are assuming to have an infinite system as only then true free-diffusion makes sense.

For large t the logarithmic term in the inequality can be neglected, and one obtains the classical *ballistic spreading* of the infection wave [107, 257]

$$x \leq 2v_0 t, \tag{3.22}$$

with constant velocity that increases monotonically with the rates [32]

$$v_0 = \sqrt{(\beta - \mu)Q}. \tag{3.23}$$

This shows that, if $p_x(t)$ is the solution for the standard diffusion problem, then there is an upper bound for the infection front propagating at constant speed.

The crucial assumption to obtain (3.22) is the "Gaussianity" of $p_x(t)$ (see also the discussion in [186]). This is justifiable in the normal diffusive case, when movements have a finite scale and are in that sense "small". Their trajectory, being a sum of many of those small displacements, is Gaussian by virtue of the central limit theorem [103]. However, one of the most prominent features of human mobility is a violation of this principle.

To model the fast multi-scale human mobility we now introduce long-range connections in the transport operator (3.19). We expect that adding such links will, by drastically reducing the topological distance between nodes, change substantially the dynamical properties of spreading processes. Instead of (3.19) we consider the general expression for the arbitrary topology (3.10), where the intensity of each rate connecting distant nodes keeps the information about the geographic location of the nodes. The Gaussian form of $p_x(t)$ needs to be replaced with a (α-stable) Lévy distribution [163], which plays the same role in the generalized central limit theorem, see Section 2.2. An important characteristic of these distributions is the power-law decay for large displacements [164]

$$p_x(t) \sim \frac{\Lambda t}{|x|^{1+\alpha}}, \tag{3.24}$$

where Λ includes the corresponding anomalous diffusion coefficient, and $\alpha \in (0, 2)$ characterizes the class of stable distribution. Plugging this into (3.16), with $c \leq \rho_x(t)$, we find

$$c \lesssim c_0 e^{(\beta-\mu)t} \frac{\Lambda t}{|x|^{1+\alpha}}.$$

Assuming the infection started on the origin, the diameter of the level set (3.17) is given by twice of above bound, due to (reflection) Z_2 symmetry. Hence the diameter $\omega(t)$ of the infected region, i.e. the prevalence of the epidemic in the geographical space, can be estimated as

$$\omega(t) \leq 2 \left(\frac{c_0}{c} \Lambda t \right)^{\frac{1}{1+\alpha}} e^{\frac{\beta-\mu}{1+\alpha} t}. \tag{3.25}$$

We see that the infected region grows exponentially fast, contrary to the ballistic growth (3.22) found for bounded jumps, which is a consequence of our power-law assumption on $p_x(t)$. In fact, linear growth of the infection front is only possible in systems where $p_x(t)$ decays as a stretched exponential [186].

In the next Section we demonstrate that, the power law (3.24) arises in our system whenever long-range connections are present. One might argue that, the assumed power-law decay in the transition rates is rather specific and far off the measured travel rates. In order to overcome the this problem, we model our ignorance about the actual travel rates with chance.

3.3. Spreading in random networks

3.3.1. The effective medium for scale-free mobility rates

Consider now the case, that the rates are random variables that define an ensemble of random networks. We assume that, and this is crucial for EMT to work [158], the rates of different links are statistically independent of each other. Hence $Q_{xy} = Q_{yx}$, but both are independent from Q_{xz}. We will assume that the rates decay like a power-law in the distance over $\mathbb{Z}$. In particular, we assume

$$Q_{xy} = Z_{xy}|x-y|^{-1-\alpha}, \tag{3.26}$$

where Z_{xy} are i.i.d. random variables chosen for each link. In practice we will sample the Z_{xy} from a uniform distribution, although in principle any distribution with a finite mean is viable [256]. For example, a distribution for Z_{xy} that mimics a more realistic human mobility pattern would take into account the scarceness of the long-range connections (flights), and the abundance of the short-range connections (cars, busses, underground trains and bicycles).

Using EMT, we can replace the ensemble of the random graphs with a deterministic graph that exhibits the *same* qualitative transport behavior, and that enables us to make quantitative predictions on the propagation speed. Hence, EMT enables us to find an average transport operator $\widetilde{\mathbf{\Omega}}$ of the random $\{\mathbf{\Omega}\}$ in such a way that the average transport fluxes remain unchanged. The graph corresponding to $\widetilde{\Omega}_{xy}$ with deterministic transition rates $\widetilde{Q}_{xy}$ defines the effective medium. The EMT prescription requires that $\widetilde{\Omega}_{xy}$ is associated to a graph with every edge (x, y) that could be present in the random networks, which are fully connected in our case. In addition, the spatial scaling of the random transition rates (3.26) has to be respected so that $\widetilde{Q}_{xy}$ decays with the same power law. The self-consistent equation (3.6) for the conductance $G_{xy} = \mathcal{N}Q_{xy}$, that in the thermodynamic limit leads to (3.9), in our special case yields

$$\widetilde{Q}_{xy} = \overline{Z}|x-y|^{-1-\alpha}, \tag{3.27}$$

where $\overline{Z} = \overline{Z}_{xy}$ for each edge (x, y), is the disorder average over the ensemble of the random networks, i.e. over the distribution of the Z_{xy} in our case.

In Appendix A we show that, using the transition rates (3.27) in the master equation (3.11), we get the superdiffusive profile (3.24) with a precise expression for the anomalous diffusion constant Λ. In particular, we find

$$p_x(t) \sim \frac{\alpha \overline{Z} t}{|x|^{1+\alpha}}. \tag{3.28}$$

Then, as long as the mean transition rate is finite, the effective medium (3.28) is exactly the deterministic long-range system (3.24) with $\Lambda = \alpha\overline{Z}$. Recently, it was proven under certain regularity conditions that this is the correct self-averaging limit of the random walk in the random network [65]. This result also defines the missing constant $\Lambda = \alpha\overline{Z}$ in (3.25) and provides the effective medium estimate for the infection diameter

$$\omega(t) \leq 2 \left(\frac{c_0}{c} \alpha \overline{Z} t \right)^{\frac{1}{1+\alpha}} e^{\frac{\beta-\mu}{1+\alpha} t}. \tag{3.29}$$

By inverting the last equation, we can also define a *generalized velocity* of the infection wave front

$$v(t) = \omega(t) t^{-\frac{1}{1+\alpha}} e^{-\frac{\beta-\mu}{1+\alpha} t}. \tag{3.30}$$

The upper bound for the diameter (3.29), determined by the effective medium approximation together with the Feynman-Kac argument, gives also an upper bound for $v(t)$, the threshold velocity

$$v_c = 2 \left(\alpha \overline{Z} \frac{c_0}{c} \right)^{1/1+\alpha}. \tag{3.31}$$

Note that, this quantity is independent on the transmission and recovery rates β and μ. Besides, this implies that the topology encoded in the universal exponent α plays a key role in identifying the threshold for a global outbreak in the metapopulation.

Although the exponential growth of the infected population is well known in the literature [186, 89, 47], EMT provides a new way to *analytically* compute quantities such as the speed of the infection or the infected diameter. Importantly, the method presented here is not limited to the considered topology nor to the particular form of the transition rates (3.26). For more general topologies and different scaling relations, a different effective medium has to be chosen. In the following Section we validate our prediction (3.29) with numerical simulations of epidemic spreading on random metapopulations.

3.3.2. Epidemic prevalence in random metapopulations

To validate our theory, we consider a ring of N subpopulations with transition rates defined by (3.26). As mentioned above, the actual distribution of $Z_{x,y}$ does not matter,

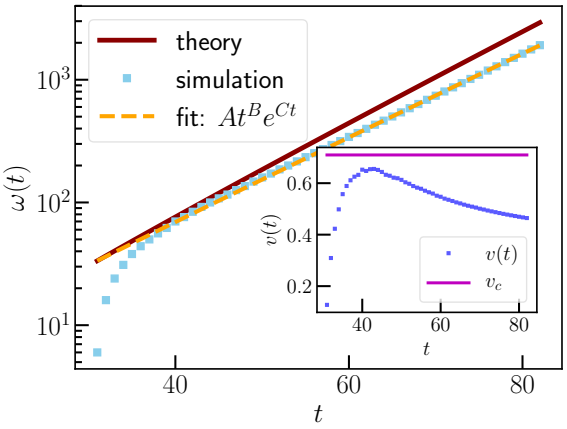


Figure 3.3: Diameter of the infected population obtained from the simulation of the metapopulation model (light-blue dots) and the EMT prediction (dark solid line), given by the upper bound of (3.29). Results are for the SI reaction in $N = 4000$ subpopulations with transmission rate $\beta = 0.2$ and Lévy exponent $\alpha = 1.5$. The numerical fit of the simulation before saturation is shown by the dashed orange line, yielding $C_{fit} = 0.076$ and consequently $\alpha_{fit} = 1.622$. Inset: generalized velocity (3.30) (blue dots) and the upper bound (3.31) (violet solid line).

hence we sampled them uniformly from the interval $[0, 1]$. Therefore $\overline{Z} = 0.5$ in our simulations.

The metric of the ring (3.18) determines the upper triangle Ω_{xy}, while the lower triangle is obtained by the symmetry condition of $\Omega_{xy} = \Omega_{yx}$. The diagonal elements are the negative sum of all other elements in the respective row. For each random realization of Ω_{xy}, equation (3.13) is integrated numerically using a fifth order Runge-Kutta method to obtain a collection $\{\rho_x(t|\Omega_{xy})\}$. Then we compute the average $\rho_x(t) = \overline{\rho_x(t|\Omega)}$ over 50 realizations of Ω_{xy}. Finally, given the infection threshold $c = 0.1$, we compute the infection diameter via (3.17).

To begin we consider a simple SI reaction scheme with transmission rate $\beta = 0.2$ and scaling exponent $\alpha = 1.5$ in $N = 4000$ subpopulations with initial concentration of infected at the origin $c_0 = 10^{-2}$. A comparison of $\omega(t)$ with the upper bound in (3.29) is given in Figure 3.3 where we also show in the inset the generalized velocity (3.30). The numerical data respects the bound nicely. However, the time series for $\omega(t)$ is truncated very early, because it saturates soon after (then the infection has reached the other side of the ring). We find that, also the generalized velocity $v(t)$ is well bounded by the value (3.31) in this case.

Since the numerical diameter $\omega(t)$ shows a nice exponential growth pattern like in (3.29), we can extract some of the parameters from the exponential fit

$$\omega(t) = At^B e^{Ct}. \tag{3.32}$$

Comparison with (3.29) would give measured values for α, $(\beta - \mu)$ and the anomalous diffusion coefficient $\overline{Z}$. This may however be a hard task, because the non-linear term

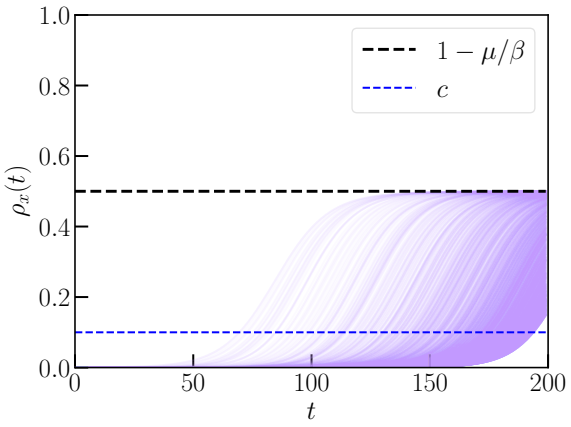


Figure 3.4: Prevalence curves (violet) for the SIS reaction with transmission and recovery rates $\beta = 0.2$ and $\mu = 0.1$ of the $N = 8000$ fully connected subpopulations with Lévy exponent $\alpha = 1.5$. The SIS stationary state for each subpopulation $\rho_x(\infty) = (\beta - \mu)/\beta$, is marked by the black dashed line while the concentration threshold c that defines the infection outbreak in each population is marked in blue. The time gap between the outbreaks of the first and last subpopulation infected is 124 time steps, and the absolute global infection time is 193 time steps.

t^B is not easy to detect in the exponential fit. The diameter's *growth rate*

$$C = \frac{\beta - \mu}{1 + \alpha}, \tag{3.33}$$

on the other hand is easy to obtain, as it can also be measured from the slope of the tail of $\ln \omega(t)$. In our simulations of the SI metapopulation model we obtain $C_{fit} = 0.076$, which gives $\alpha_{fit} = 1.622$. This is a reasonably close value to the Lévy exponent $\alpha = 1.5$ used for generating the graph realizations in the first place.

We now consider the SIS model in a larger network of $N = 8000$ subpopulations with transmission and recovery rates $\beta = 0.2$ and $\mu = 0.1$, resulting in a basic reproductive number $\mathcal{R}_0 = 2$. The correct time frame to assess $\omega(t)$ is visible in a prevalence plot, see Figure 3.4. The curves $\rho_x(t)$ for each x are plotted against time and the stationary value $\rho_x(\infty) = 1 - \mu/\beta$ as well as the time when $\rho_x(t) > c$, can be read in the same plot. For the estimation of the Lévy exponent at this reproductive number we find $C = 0.041$, which results in $\alpha_{fit} = 1.454$, which is off by only 3% of the theoretical value $\alpha_{the} = 1.5$.

Varying the basic reproductive number and measuring the growth rate C or the Lévy exponent α, respectively, leads to good coincidence between theory and numerics, see Figure 3.5 (a). When the theoretical Lévy exponent α is varied and the growth rate is measured, the agreement appears much worse, see Figure 3.5 (b). This mismatch is easily explained as a finite size effect, as we show in Figure 3.5 (c). There, we plot the difference $\Delta C = |C_{the} - C_{fit}|$ between the measured growth rate C_{fit} and its theoretical prediction C_{the} from (3.33) in a double logarithmic fashion against the system size N. The figure shows that the error decays at least as a power law and will vanish in the thermodynamic limit $N \to \infty$. As for the SI results, due to the extreme long-range connections, $\rho_x(t)$ saturates very quickly. This leads to very short time frame in which $\omega(t)$ grows exponentially, making a correct estimation of C very difficult. The effect

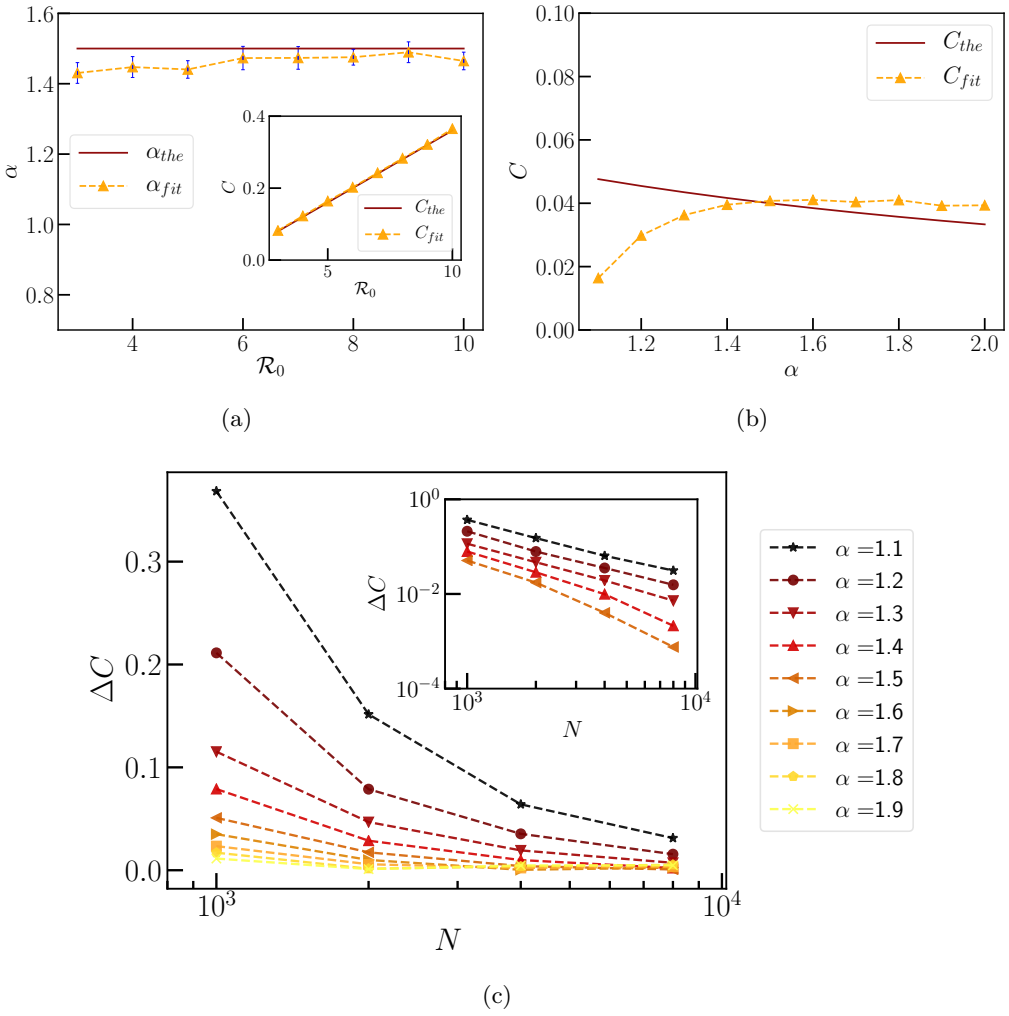


Figure 3.5: (a) The extrapolated value of the Lévy exponent α with the corresponding error (blue bars) evaluated from the error propagation of the numerical fit error in C, shown in the inset, as a function of the basic reproductive number $\mathcal{R}_0 = \beta/\mu$ for the given theoretical value $\alpha_{the} = 1.5$. (b) Theoretical growth rate C_{the} and the simulation fitted value C_{fit} for the SIS reaction with transmission and recovery rates $\beta = 0.2$ and $\mu = 0.1$, in $N = 8000$ subpopulations as a function of the Lévy exponent $\alpha \in (1, 2]$. (c) Absolute value of the difference $\Delta C = |C_{the} - C_{fit}|$ between the theoretical $C_{the} = (\beta - \mu)/(1 + \alpha)$ and the simulation fit value C_{fit} for the SIS reaction as a function of the subpopulations number N with $\beta = 0.2$ and $\mu = 0.1$. Different lines are for different Lévy exponents from dark to light in the range $\alpha \in (1, 2)$. Inset: close-up in doubly logarithmic scale for $\alpha \in (1, 1.5)$. For larger values of α, the error fluctuates around 0.005 which is the numerically attainable accuracy.

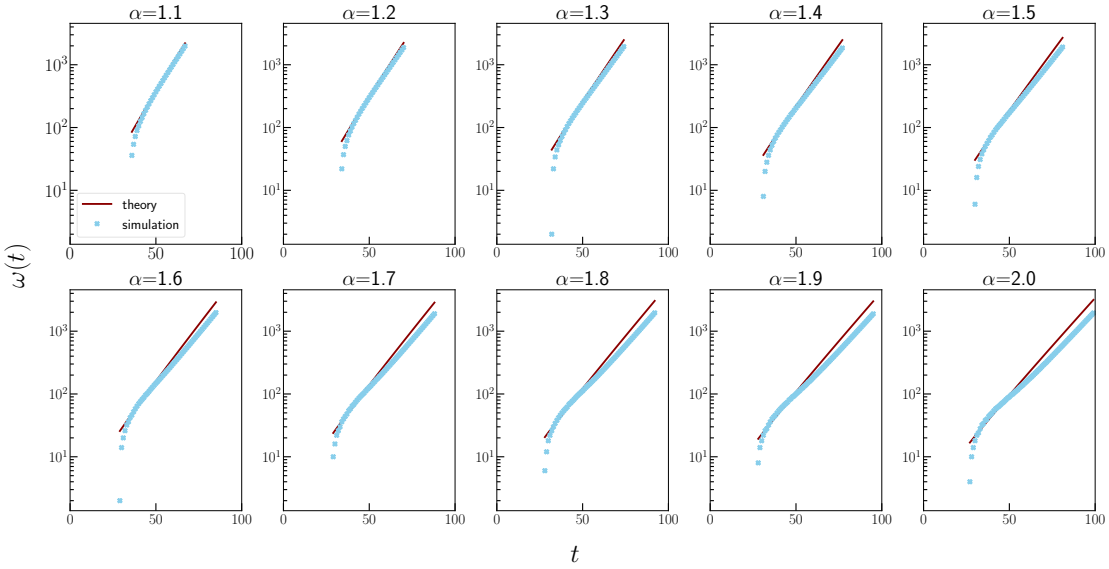


Figure 3.6: Diameter of the infected population obtained from the simulations (light-blue dots) of the SIS reaction in $N = 8000$ subpopulations with transmission and recovery rates $\beta = 0.2$ and $\mu = 0.1$, and the theoretical prediction (dark solid line) given by EMT for various Lévy exponent α.

becomes worse as α decreases, which also explains the slightly worse agreement for small α in Figure 3.5 (a).

As a final remark, we note that in Figure 3.5 (b) we found $C_{fit} > C_{the}$ in that data range, which should not be possible as (3.29) represents an upper bound. An overview of the agreement with the theoretical bound for the probed range in α is shown in Figure 3.6. We find that for some of these large values of α the numerical data overestimate the EMT bound. This, however, only happens in an intermediate time regime and not in the long time limit, in which we derived (3.29). In fact, we find that the upper bound is respected in the long-time limit for all values of α. The predictions given by EMT are rigorously valid in the thermodynamic limit $N \to \infty$, when the infection propagates indefinitely and saturation is never reached.

In this Chapter, we have presented a new analytical tool for studying reaction-diffusion problems in random media. We wanted to advocate the use of EMT that provides a deterministic representative for an ensemble of random networks. Together with the Feynman-Kac argument, we provided an upper bound for the infection front using the SIS model. This way we demonstrated that EMT is still relevant even beyond the short-range connection paradigm to study epidemic spreading in random networks. In order to model real human mobility, we used a metapopulation model with random long-range connections lacking a definite spatial scale. Due to the presence of long-range connections we found the exponential growth of the infection diameter. The growth rate

depends on both the infection and recovery rates β and μ, as well as on the topology encoded in the Lévy exponent α of the statistical decay of the mobility rates (3.26). Long-range connections with a "weak" power law, i.e. $\alpha \geq 2$, would have eventually lead to a ballistic growth of the infection front [132]. EMT is known to nicely reproduce the transport behavior of a random system *provided* it is far away from the percolation threshold. The random networks we treated here are very well connected (in fact all connections besides self-loops are present) due to the presence of the long-range links. Therefore, they are always far from percolation threshold, which partly explains the success of our approach.

It might seem that the model studied here lacks realism as we considered a one dimensional array of subpopulations. However, the long-range connections dominating the dynamics make our results quite general and applicable to other underlying topologies. In particular, a measurement of C, available from a simple linear fit of the logarithm of the infection diameter, can give a rough estimate of the exponent α of the transition rates. The numerical fit estimation of the EMT parameters becomes reliable for the SIS model only for sufficiently large networks as finite size effects are quite severe in this case. We believe that modifications of long-range EMT will remove some restrictions that we imposed in our treatment. For example, modifications of EMT are necessary to deal with random networks where the transition rates Q_{xy} are not symmetric and variable subpopulation sizes can be taken into account. An extension of our argument to d dimensions is also possible without major change and would only lead to a different growth rate of $C = (\beta - \mu)/(d + \alpha)$. Internal dynamics on the nodes (like commuting agents) could be considered by replacing the subpopulations by small networks themselves. To overcome the strong finite size effects that we encountered in our simulations, one alternative approach is to consider a finite size effective medium instead of an infinite one, as we did here for simplicity. Finally, when it is possible to deal with *correlated* links, more realistic models than a regular lattice can be included such as very heterogeneous topologies observed in real human-mobility networks [21].

4

The Hidden Geometry of Spreading Processes

"*I have no idea where this will lead us, but I have a definite feeling it will be a place both wonderful and strange.*"

–Dale Cooper

Contents

The forecast and control of emergent diseases spreading at the global scale has become particularly important in recent years because of the exponential growth in both the structure and velocity of transportation means. In this Chapter, we introduce a network-based measure that generalizes the concept of geodesic distance (2.2) and that provides fundamental insights into the dynamics of disease transmission as well as an efficient numerical estimation of the infection arrival time. We compare this *effective*

distance (ED) with the numerical estimate of the arrival times using the metapopulation model, see Section 2.3.4. A series of papers have already been devoted to this problem [43, 120, 121, 46]. However, most of them are based on the assumption that a single dominant path, associated to maximal traffic probability, is sufficient to estimate accurately the arrival time of a diffusive process. While this is partially true in some specific cases, when a single path between each pair fo nodes is available, in the general scenario this assumption can give bad estimates for the arrival time. Effective distances in the *dominant-path* approach can be defined, for both directed and undirected networks, as the geodesic graph distance (2.2) of a weighted graph with edge weights given by the first moment of a distribution known from extreme events statistics [131], which depends only on the network topology and on the transmission and recovery rates. This approach has the disadvantage that it can significantly overestimate the numerical infection arrival time [120, 82]. In addition, in situations where multiple equiprobable paths exist between node pairs, as in regular lattices, this approach breaks down. A more realistic scenario takes into account all possible propagation routes [121] yielding a *multiple-path* ED, which is supposedly the best possible estimate of infection arrival times. Unfortunately, the computation becomes infeasible as the number of paths between two nodes grows exponentially with the size of the network. The lack of a practical computational approach leads back to considering only the dominant path. We introduce a *random-walk* approach that generalizes the multiple-path ED and that can be used as a computationally feasible alternative with a clear interpretation, paving the way for future studies on information transfer in complex networks.

The Chapter is organized as follows. In Section 4.1 we describe the metapopulation model on the global mobility network of air-traffic. This particular dataset, as provided by the Official Airline Guide [1], is used to simulate real-world scenarios of global pandemics with different infection seeds. Effective distances in the dominant-path, multiple-path and random-walk approaches are derived in Section 4.2. We first define the dominant-path effective distance by considering node-independent exit rates and then generalize the multiple-path approach to random walks. Finally, in in Section 4.3 we perform numerical experiments of epidemic spreading in the global mobility network, the air-traffic network in the United States and three additional artificial networks. The results presented in this Chapter are discussed in [148].

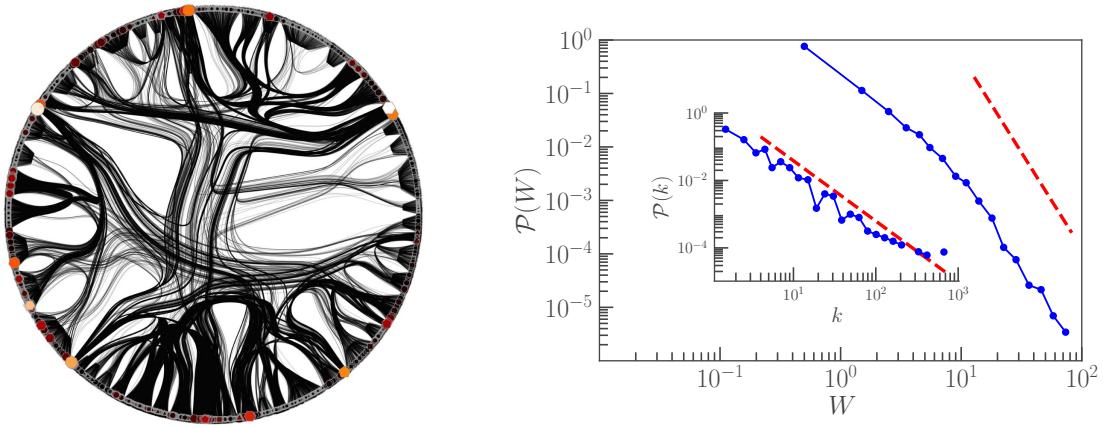
4.1. The global mobility network

As a proxy for global pandemics we perform numerical experiments with the metapopulation model using the global mobility network (GMN). The GMN is constructed from a dataset provided by Official Airline Guide [1]. The dataset includes a total of $N = 3865$ airports and the number of seats on scheduled commercial flights between pairs of airports over the three-year period 2004-2006, see Figure 4.1. Assuming that the number

Figure 4.1: The global mobility network (GMN) of air-traffic as provided from the Official Airline Guide (OAG Ltd.) [1]. Each edge corresponds to a scheduled commercial flight over the three-year period 2004-2006, with gradient scaling from dark to light-blue according to the available number of seats.

of seats on scheduled commercial flights is on average proportional to the number of passengers traveling, the data can be represented as a weighted network with adjacency matrix W_{ij} giving the total traffic per day between airport i and Airport j. To each airport j (a node in the metapopulation) is associated a subpopulation of size N_j so that $W_{ij} = Q_{ij}N_i$, where Q_{ij} is the transition rate from i to j. Although the network is directed, the degree of asymmetry in the weighted adjacency matrix W_{ij} is extremely small. This particular feature is well confirmed also in similar empirical datasets of air-traffic at the global scale [22, 46]. We estimate quantitatively the degree of asymmetry by looking at both the topological and weights asymmetry. The former is quantified by the average number of non-zero elements ϵ in the corresponding unweighted adjacency matrix $A_{ij} = \chi(W_{ij})$, where χ is the step function equal to one for positive arguments and vanishing otherwise. Instead the weight asymmetry ϵ_w is defined as the normalized net difference between travel fluxes in each route and the corresponding reversed travel. We find $\epsilon = 2 \cdot 10^{-3}$ and $\epsilon_w = 3.1 \cdot 10^{-9}$. Thus, we redefine the a symmetric adjacency matrix as $W_{ij} = (W_{ij} + W_{ji})/2$. Symmetrizing the weights also assures us that the subpopulation sizes are conserved quantities, see (2.82).

In principle, the matrix W_{ij} is provided by traffic data and N_i by census data such that the rates Q_{ij}, for $i \neq j$, can be computed inverting $W_{ij} = Q_{ij}N_i$. However, although it is straightforward to measure W_{ij}, assessing the effective population is more subtle. The number of individuals that effectively participate in the dispersal N_i is not necessarily the same as the population data provided by census. As an alternative, see Section



(a) (b)

Figure 4.2: (a) Circular representation of the GMN with nodes color and size scaling according to the corresponding strength $s_i = \sum_j W_{ij}$, from black to white. (b) Weights distribution $\mathcal{P}(W) \sim W^{-\delta}$ with scaling exponent $\delta = 3.60 \pm 0.14$. Inset: (unweighted) topological degree distribution $\mathcal{P}(k) \sim k^{-\gamma}$ with scaling exponent $\gamma = 1.79 \pm 0.10$. Scaling exponents are obtained using the method described in [69].

2.3.4 we also assume that the exit rate $q_i = \sum_{l \neq i} Q_{il}$ is independent of the node i and reformulate the diffusion contribution in terms of the transition matrix (2.88) and a constant diffusion rate (2.87). The latter is given by the ratio $\alpha = \mathcal{W}/\mathcal{N}$, between the total flux (per unit time) $\mathcal{W} = \sum_{ij} W_{ij}$ and the total population $\mathcal{N} = \sum_i N_i$. In terms of α we can remove the dependence on the subpopulation size N_i from the reaction-diffusion equations and use (2.89). The symmetrized GMN consists of $N = 3865$ nodes (airports) and $E = 26691$ undirected edges (routes), with very broad degree and weight distributions, see Figure 4.2. For the network diameter and global clustering we find $D = 16$ (connecting *Stuart Island* to *Narsaq Kujalleq Heliport*) and $\langle C \rangle = 0.26 \pm 0.01$, respectively. A summary of the relevant statistical properties of the GMN, along with the additional networks used in for the numerical analysis in Section 4.3, is given in Table 4.1.

4.2. Effective distances

The fundamental metric in networks is the geodesic distance, i.e. the shortest-path length over all paths $\{\Gamma_{ij}\}$ connecting node i to node j. For weighted networks, where the weights are positive numbers identifying the carrying capacity of a certain route, each edge $(k, l) \in \Gamma_{ij}$ contributes to the total length with its reciprocal weight. This

	N	E	D	$\langle C \rangle$	$\langle k \rangle$	$\langle k^2 \rangle$
GMN	3865	26691	16	0.26	13.81	1032.81
USA	500	2980	7	0.35	11.92	641.12
ER	500	2980	4	0.05	11.92	153.47
BA	500	2485	4	0.05	9.94	181.88
RL	500	1000	30	0	4	16

Table 4.1: Statistical properties of the networks used in the numerics: the global mobility network (GMN) the air-traffic network of the United States of America (USA), its edge-randomized version with boolean weights (ER), an unweighted Barabási-Albert network (BA) with $m = 5$ new edges per timestep and an unweighted regular lattice (RL). The different quantities are: the number of nodes N, the number of edges E, the diameter D, the global clustering $\langle C \rangle$, the first moment $\langle k \rangle$ and the second moment $\langle k^2 \rangle$ of the degree distribution.

generalizes the standard definition for unweighted networks (2.2) as

$$D_{ij} = \min_{\{\Gamma_{ij}\}} \sum_{(k,l)\in\Gamma_{ij}} \frac{1}{W_{kl}}. \tag{4.1}$$

The reciprocal of the weights is used consistently with the fact that a higher flux of passengers along an edge reduces the distance between the respective nodes.

Starting from the heuristic definition (4.1), it is possible to extend the notion of distance by replacing the weights with an *effective function* of the weights $f(W_{kl})$ that quantitatively reproduces the distance as measured by the arrival time of spreading processes unfolding on a given network. A naive but effective ansatz was proposed in [46]. The authors define the ED as

$$D_{ij}^{\text{eff}} = \min_{\{\Gamma_{ij}\}} \sum_{(k,l)\in\Gamma_{ij}} (1 - \ln P_{kl}), \tag{4.2}$$

where $P_{kl} = W_{kl}/\sum_m W_{km}$ is the transition probability to navigate the graph via random walks. The choice for the logarithm is motivated by the authors simply by requiring the additivity of the distances (4.2), consistently with multiplying the corresponding probabilities. Although this ED is able to reproduce accurately the infection arrival time in the GMN, its interpretation remains quite obscure. Indeed, several questions are in order on the specific form (4.2). Besides serving the function of multiplying the probabilities when distances are added, there is not a transparent explanation for the logarithm. Furthermore the choice for the constant term equal to unity in its definition is quite arbitrary and it is not clear *a priori* why such expression would correctly quantify the arrival times of reaction-diffusion processes in arbitrary networks. Finally, the definition (4.2) suffers from the great limitation of considering a single (probability-dominant) path, neglecting all other routes for information transfer.

4.2.1. Dominant path

An alternative and refined approach to derive analytically using a detailed kinetic description of the spreading process is outlined in the following. Interestingly, (4.2) can be obtained as a special case of a more general quantity. To show this we extend and generalize the Markovian description presented in [121] and define the ED using a dominant-path approach, by requiring the maximization of the travel probability between adjacent subpopulations. With respect to the derivation given in [121], we will take a step further and assume that the exit rate of each node is independent on that location to obtain an expression with the same form of (4.2), but with an additional tunable constant that depends on the model parameters.

Let us first consider the simpler metapopulation model with SI reaction and only two subpopulations i and j, with initial condition $I_i(0) = 1$. In a Markovian description at every time step Δt, each of the I_i infected individuals in subpopulation i has a probability $p = Q_{ij}\Delta t$, to jump to subpopulation j, where $Q_{ij} = W_{ij}/N_i$ are the transition rates and W_{ij} the flux of passengers per time step in the (directed) edge (i, j). The probability that the first infected arrives from i to j after n time steps $n_j = n\Delta t$, the infection arrival time to j, is given by [121]

$$P(n_j = n\Delta t) = \left(1 - (1-p)^{I_i(n\Delta t)}\right) \prod_{k=1}^{n-1} (1-p)^{I_i(k\Delta t)}. \tag{4.3}$$

In this equation $(1-p)^{I_i(n\Delta t)}$ is the probability that no infected individual in i moves to j at exactly time $n\Delta t$ and in the product all the probabilities of not jumping at all times before $n\Delta t$ are multiplied. In the limiting case of low mobility $p \ll 1$, we can approximate $(1-p)^k \approx e^{-pk}$ and $\left(1 - (1-p)^k\right) \approx pk$. Then (4.3) can be rewritten as

$$P(n_j = n\Delta t) \approx pI_i(n\Delta t) \exp\left(-p \sum_{k=1}^{n-1} I_i(k\Delta t)\right). \tag{4.4}$$

The continuous-time limit

$$\Delta t \sum_{k=1}^{n-1} I_i(k\Delta t) \longrightarrow \int_0^{t=n\Delta t} d\tau \; I_i(\tau), \tag{4.5}$$

yields the probability density function for the first arrival of the infection at time $t = n\Delta t$

$$\mathcal{P}(n_j = t) \approx Q_{ij} I_i(t) \exp\left(-Q_{ij} \int_0^t I_i(\tau) d\tau\right). \tag{4.6}$$

Finally, assuming the early stage of the epidemic where the linearization of (2.67) gives

(2.72) for $\mu = 0$, using the initial condition $I_i(0) = 1$ yields

$$\begin{aligned}\mathcal{P}(n_j = t) &\approx Q_{ij} \exp\left(\beta t - \frac{Q_{ij}}{\beta} e^{\beta t}\right) \\ &\approx \beta \exp\left(\beta t + \ln \frac{Q_{ij}}{\beta} - \exp\left(\beta t + \ln \frac{Q_{ij}}{\beta}\right)\right). \end{aligned} \tag{4.7}$$

This probability density function is known in the theory of extreme events as the Gumbel minimum[1] distribution [131]

$$\mathcal{P}(t) = \frac{1}{a} \exp\left(\frac{1}{a}(t-b) - e^{\left(\frac{1}{a}(t-b)\right)}\right), \tag{4.8}$$

where a and b are the scale and location parameter respectively.

The first moment of (4.7) is

$$\langle n_j \rangle_i = \int_{t_i=0}^{t_j=n_j} dt\ t\mathcal{P}(t) = \frac{1}{\beta}\left(-\ln \frac{Q_{ij}}{\beta} - \gamma_e\right), \tag{4.9}$$

where $\gamma_e \approx 0.577$ is the Euler-Mascheroni constant and the average $\langle \ldots \rangle_i$ is taken over the arrival times of infected individuals with initial condition $I(t_i) = 1$. This can be rewritten, assuming as in Section 2.3.4 that the exit rates $q_i = \sum_{l \neq i} Q_{il} = \alpha$ are node independent, as

$$\beta \langle n_j \rangle_i = \lambda - \ln P_{ij}, \tag{4.10}$$

where $\lambda = \ln \beta/\alpha - \gamma_e$ is constant and P_{ij} is the transition matrix (2.88).

The previous derivation based on the Markovian assumption for the diffusion of agents, can be immediately generalized to arbitrary metapopulation networks with permanent immunization described by the SIR reaction scheme with recovery rate μ. This yields the *dominant-path* ED

$$D_{ij}^{\mathrm{DP}}(\lambda) = \min_{\{\Gamma_{ij}\}} \sum_{(k,l)\in\Gamma_{ij}} (\lambda - \ln P_{kl}), \tag{4.11}$$

where

$$\lambda = \ln \frac{\beta - \mu}{\alpha} - \gamma_e. \tag{4.12}$$

From (4.11) we can recover the ansatz for the ED (4.2) simply by setting $\lambda = 1$. However, in the dominant-path ED (4.11), the definition of λ as a function of the epidemic and mobility parameters gives the optimized edge weight that should contribute into the

[1] The Gumbel maximum distribution is obtained with the substitution $(t, b) \to (-t, -b)$.

minimization condition over all paths connecting source and target. On the computational side, one can obtain the full matrix D_{ij}^{DP} using the Dijkstra algorithm [91] in a time $\mathcal{O}(NE + N^2 \log N)$, where E and N are the graph size and order, respectively.

The most important limitation of (4.11) is that only the path that minimizes the topological length and at same time maximizes the associated probability is considered. It turns out that because of this limitation, the effective infection arrival time $D_{ij}^{\mathrm{DP}}(\lambda)/v^{\mathrm{eff}}$, where $v^{\mathrm{eff}} \approx \beta - \mu$ is the linearized effective speed of the infection [46], is overestimated with respect to the numerical arrival times obtained from direct simulations [82, 120].

4.2.2. Multiple paths

The correct approach is to consider the multiplicity of transmission routes. The framework to include all possible paths of transmission was developed in [121]. In the following we review and extend their derivation to obtain an expression for the ED that can easily be generalized to random walks. For simplicity we start by considering only two distinct paths Γ and Γ' connecting the same pair of nodes. The two-paths ED D_{ij}^{2P}, that generalizes the dominant-path ED (4.11) satisfies the equation

$$e^{-D_{ij}^{\mathrm{2P}}} = e^{-D_{ij}^{\Gamma}} + e^{-D_{ij}^{\Gamma'}}, \tag{4.13}$$

where

$$D_{ij}^{\Gamma}(\lambda) = \sum_{(k,l)\in\Gamma_{ij}} (\lambda - \ln P_{kl}) = -\ln\left(\prod_{(k,l)\in\Gamma_{ij}} e^{-\lambda} P_{kl}\right) \tag{4.14}$$

is the ED, which is the mean of a Gumbel distribution, associated to a path Γ_{ij} of arbitrary length connecting node i to node j. Relation (4.13) can be easily generalized to an arbitrary number of paths connecting the same pair fo nodes, as

$$\exp\left(-D_{ij}^{\mathrm{MP}}(\lambda)\right) = \sum_{\{\Gamma_{ij}\}} \exp\left(-D_{ij}^{\Gamma}(\lambda)\right). \tag{4.15}$$

The previous equation defines the *multiple-path* ED

$$D_{ij}^{\mathrm{MP}}(\lambda) = -\ln\left(\sum_{\{\Gamma_{ij}\}} e^{-\lambda n(\Gamma)} P(\Gamma)\right), \tag{4.16}$$

where the total probability associated to the path Γ_{ij} of length $n(\Gamma_{ij}) = |\Gamma_{ij}|$ is

$$P(\Gamma_{ij}) = \prod_{(k,l)\in\Gamma_{ij}} P_{kl}. \tag{4.17}$$

An analogous expression can be obtained by grouping all probabilities associated to paths of same length into the quantity $F_{ij}(n) = \sum_{|\Gamma|=n} P(\Gamma_{ij})$. Then we can replace the sum over all paths connecting i to j in (4.16) with a sum over the allowed path lengths $n = |\Gamma| \in [1, n_{max}]$, to get

$$D_{ij}^{\mathrm{MP}}(\lambda) = -\ln\left(\sum_{n=1}^{n_{max}} e^{-\lambda n} F_{ij}(n)\right), \tag{4.18}$$

where n_{max} is the maximum path length in the network. If instead of considering all paths in (4.16) we select the single path Γ_{ij}^{*} of length $n(\Gamma_{ij}^{*}) = |\Gamma_{ij}^{*}|$ that is associated to the dominant contribution, i.e. the path that maximize its associated probability and minimizes the topological path length, we recover the dominant-path ED (4.11), i.e.

$$D_{ij}^{\mathrm{DP}}(\lambda) = D_{ij}^{\Gamma^{*}}(\lambda) = -\ln\left(e^{-\lambda n(\Gamma_{ij}^{*})} P(\Gamma_{ij}^{*})\right). \tag{4.19}$$

Although the multiple-path ED gives the most accurate estimate of the infection arrival time, as it counts the most probable route as well as all possible alternative transmission routes, it is computationally not tractable. In fact since the total number of paths between i and j can scale as $\mathcal{O}(N!)$, the measure D_{ij}^{MP} becomes useless for large networks [260]. A trade-off between performance and accuracy can be achived by restricting the path search algorithm to a maximum path length or more elegantly, as we show in the next Section, by relaxing the assumption of direct propagation.

4.2.3. Random walks

Both measures introduced in the previous sections D_{ij}^{DP} and D_{ij}^{MP} rely on the fact that the epidemic will spread along paths, i.e. routes that do not ever cross. Here we follow a different approach and introduce an ED that includes all possible random walks from source to target. Relaxing the assumption of directed spread is equivalent to effectively erasing the memory from the system at each time step. This is achieved by including in (4.16) all walks $\{\Xi\}$ that, contrary to the paths $\{\Gamma\}$, allow also crossing already visited nodes. We define the *random-walk effective distance* (RWED) by generalizing (4.16) as

$$D_{ij}^{\mathrm{RW}}(\lambda) = -\ln\left(\sum_{\{\Xi_{ij}\}} e^{-\lambda n(\Xi)} P(\Xi)\right), \tag{4.20}$$

where $P(\Xi_{ij}) = \prod_{(k,l)\in\Xi_{ij}} P_{kl}$ is the total probability associated to the walk Ξ_{ij} of length $n(\Xi_{ij}) = |\Xi_{ij}|$. We note that since $\{\Xi\}$ is a bigger set than $\{\Gamma\}$, the following inequalities hold

$$D_{ij}^{\mathrm{RW}} \leq D_{ij}^{\mathrm{MP}} \leq D_{ij}^{\mathrm{DP}}. \tag{4.21}$$

As we did in the previous section for paths with probability $P(\Gamma)$, we can group the probabilities associated to walks of the same length into the quantity $H_{ij}(n) = \sum_{|\Xi|=n} H_{ij}(\Xi_{ij})$. The latter is precisely the hitting time probability (2.47) that is also defined recursively for $i \neq j$ as $H_{ij}(n) = \sum_{k \neq j} P_{ik} H_{kj}(n)$. Contrary to the multiple-path scenario, walks are unbounded and so becomes the maximum length n_{max} in (4.18). Substituting, we rewrite the RWED as

$$D_{ij}^{\mathrm{RW}}(\lambda) = -\ln\left(\sum_{n=1}^{\infty} e^{-\lambda n} H_{ij}(n)\right). \tag{4.22}$$

Remarkably, there is a immediate interpretation of the RWED in terms of the hitting-times generating functions defined in Section 2.2.2. Indeed by definition

$$D_{ij}^{\mathrm{RW}}(\lambda) = -\ln\left\langle e^{-\lambda n_{ij}}\right\rangle, \tag{4.23}$$

where n_{ij} is the random-walk hitting time to node j [154] and the average is taken over all random walks that start in i and terminate as soon as they hit node j. Comparing with the definition of cumulant generating function (2.49) of the random-walk hitting time, it is easy to see that

$$D_{ij}^{\mathrm{RW}}(\lambda) = \sum_{k=0}^{\infty} (-1)^{k+1} \frac{\lambda^k}{k!} \left\langle n_{ij}^k \right\rangle_c = -\Psi_{ij}(-\lambda), \tag{4.24}$$

where $\left\langle n^k \right\rangle_c$ are the hitting time cumulants. Hence one obtains the cumulants of the random-walk hitting time by differentiating (4.22) with respect to λ. In particular the mean-first-passage time (MFPT) from i to j is obtained as $M_{ij} = \left\langle n_{ij} \right\rangle_c = \partial_\lambda D_{ij}^{\mathrm{RW}}|_{\lambda=0}$. To compute $D_{ij}^{\mathrm{RW}}(\lambda)$ we can write $H_{ij}(n)$ in terms of powers of the sub-transition probability matrix $\mathbf{P}^{(j)}$ obtained by removing the jth row and column, as for the computation of the MFPT in Section 2.2.2. Assuming a positive λ, the expansion in a geometric series converges[2] and we obtain

$$D_{ij}^{\mathrm{RW}}(\lambda) = -\ln\left(\sum_{k \neq j} \left(\mathbf{I}^{(j)} - e^{-\lambda}\mathbf{P}^{(j)}\right)^{-1}_{ik} e^{-\lambda} p_k^{(j)}\right). \tag{4.25}$$

Here, $\mathbf{I}^{(j)}$ is the $(N-1) \times (N-1)$ sub-identity matrix and $\mathbf{p}^{(j)}$ is the jth column of $\mathbf{P}$ with jth component removed that takes into account the last step needed to reach the target j. Using (4.25), the RWED can be computed in polynomial time $\mathcal{O}(N^{3.4})$ using e.g. the Coppersmith-Winograd algorithm for matrix inversion [78], making the

[2] In order to have a converging expression (4.22), we must require $\lambda > 0$ and then the determinant of the matrix $e^{-\lambda}\mathbf{P}^{(j)}$ is always smaller than unity. By recalling the definition (4.12), the previous condition imposes an additional constraint on the model parameters, i.e. $\beta > \mu + \alpha e^{-\gamma_e} \approx \mu + 2\alpha$.

problem of parallel transmission routes feasible even for large networks.

To conclude the Section, we show how we can disentangle the dominant contribution defined by D_{ij}^{DP} in the RWED, by using a *path-integral* formulation [285] of ED. Indeed, we can rewrite the RWED (4.20) as

$$D_{ij}^{\mathrm{RW}} = -\ln\left(\sum_{\{\Xi_{ij}\}} e^{-\mathcal{A}(\Xi,\lambda)}\right), \tag{4.26}$$

where $\mathcal{A}(\Xi,\lambda)$ is interpreted as the Euclidean action defined as

$$\mathcal{A}(\Xi,\lambda) = \lambda n(\Xi) - \ln P(\Xi). \tag{4.27}$$

As previously, $n(\Xi)$ is the length of the walk Ξ and $P(\Xi)$ is the associated probability. In this picture the RWED is the the free energy functional (changed in sign) of an Euclidean field theory [212, 285]. The dominant-path ED is instead described by (4.19). Contrary to (4.26), the dominant-path ED neglects contributions of paths other than the one that maximizes the associated probability. Then, we can think of D_{ij}^{DP} as an *effective action*[3]

$$D_{ij}^{\mathrm{DP}} = \mathcal{A}^{\mathrm{eff}}(\Gamma_{ij}^{*},\lambda) = \lambda n(\Gamma_{ij}^{*}) - \ln P(\Gamma_{ij}^{*}), \tag{4.28}$$

where Γ_{ij}^{*} is the (probability) dominant path of length $n(\Gamma_{ij}^{*})$. Expanding the action around the dominant contribution as

$$\mathcal{A}(\Xi,\lambda) \approx \mathcal{A}^{\mathrm{eff}}(\Gamma^{*},\lambda) + \frac{\delta\mathcal{A}}{\delta\Xi}\Big|_{\Xi=\Gamma^{*}}\delta\Xi + \dots \tag{4.29}$$

and plugging this into (4.26) we find

$$D_{ij}^{\mathrm{RW}} = \mathcal{A}^{\mathrm{eff}}(\Gamma_{ij}^{*},\lambda) - \ln\left(\sum_{\{\Xi_{ij}\}} e^{\frac{\delta\mathcal{A}}{\delta\Xi}\Big|_{\Xi=\Gamma^{*}}\delta\Xi+\dots}\right). \tag{4.30}$$

In the previous equation $\delta\Xi$ quantifies the deviation of the walk Ξ with respect to the dominant path Γ^{*}. By comparing with the ansatz D_{ij}^{eff} defined by (4.2), we can finally identify

$$D_{ij}^{\mathrm{eff}} = \mathcal{A}^{\mathrm{eff}}(\Gamma_{ij}^{*},1), \tag{4.31}$$

where the effective action is defined by (4.28). The additional contributions in the RWED with respect to the effective action, quantified by the logarithmic term on the right-hand side of (4.30), account for the multiplicity of transmission routes. This logarithmic term quantifies the net difference between the random-walk and the dominant-path

[3] Note that our definition differs from the standard effective action of quantum field theory used to analyze the renormalizability of the theory [223].

approaches. We expect this contributions to be negligible for network topologies that are locally tree-like, where a single path connects any pair of nodes. For many real-world networks this is the case since the degree distribution behaves as a power-law and the navigation is dominated by the presence of large hubs [54]. However we expect the dominant-path approach to completely break down when the multiplicity of paths becomes relevant as in the case of random (Poissonian) networks and regular lattices where a large number of paths are equally probable. In this case the difference $|D_{ij}^{\mathrm{RW}} - D_{ij}^{\mathrm{DP}}|$ will be non-negligible and the arrival time estimates will substantially decrease when considering the random-walk approach.

4.3. Hitting times of global pandemics

For the numerical analysis we use the SIR metapopulation model defined by (2.89) for $\chi = 0$. We denote by S_j, I_j and R_j the number of individuals in subpopulation j who are susceptible, infected, and removed, respectively. The normalized quantities are $\rho_j^X = X_j/N_j$, where $X \in \{S, I, R\}$, and N_j is the constant subpopulation size, see (2.82). The key quantity we are interested in estimating is the *infection arrival time*, defined for each pair of node by the matrix

$$T_{ij} = \min_t \left\{ t \geq 0 \mid \rho_j^I(t) \geq 1/N_j \right\}_i , \tag{4.32}$$

where the subscript i implies that the process started in node i at time $t = 0$, and by definition $T_{ii} = 0$. We use T_{ij} as the benchmark of infection arrival times in real-world pandemic scenarios and compare it to the ED defined by (4.11) and (4.25).

Different definitions than (4.32) for the infection arrival time are obviously possible. Indeed, a major drawback of the metapopulation model used defined by (2.89), is that it is fully deterministic and it involves only the averaged diffusive coupling, while in reality both epidemic spreading and travel of individuals are inherently stochastic processes. Particularly, when the number of infected are small compared to the subpopulation size, i.e. when $\rho_j^I(t) \ll 1$, fluctuations can play a dominant role. A phenomenological modification of the above dynamics is based on the inclusion of a local invasion threshold η, assuming that a local epidemic can only take off when $\rho_j^I(t)$ exceeds that fixed small fraction of the subpopulation. The main idea consists in redefining a node dependent transmission rate as a threshold function $\beta(\rho_j^I(t)/\eta)$ with sigmoid shape, that triggers smoothly the infection process if the infected exceeds a fixed small fraction of the subpopulation, i.e. if $\rho_j^I(t)/\eta > 1$ for some $\eta \ll 1$ [46]. The results presented in this Chapter are robust when adopting this approach of "refined" reaction-diffusion that mimics the presence of fluctuations.

A qualitative comparison between RWED and infection arrival time is shown in Figure 4.3. There we show four different time snapshots of a pandemic in the GMN with

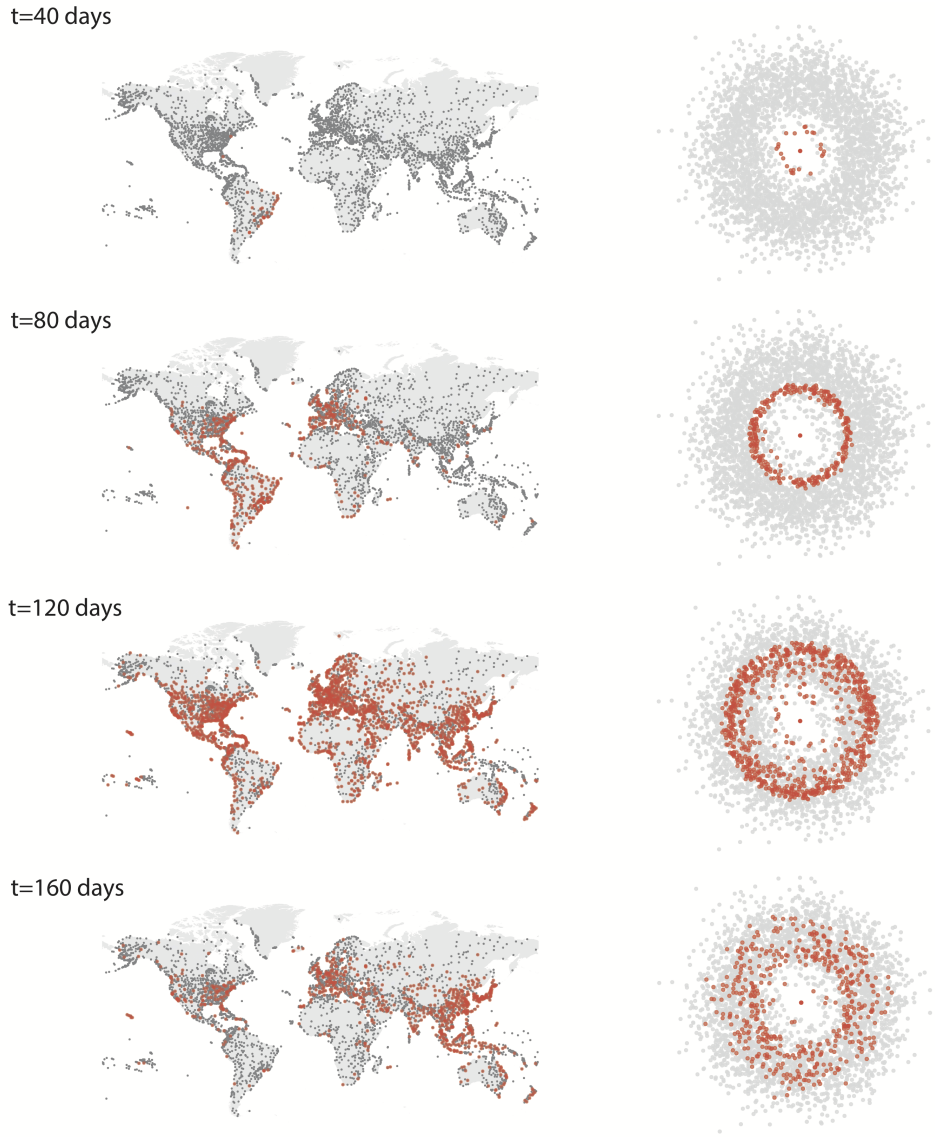


Figure 4.3: Left: Prevalence of a global pandemic with basic reproductive number $\mathcal{R}_0 = 1.5$ at four different observation times, as obtained from numerical integration of (2.89) with $\chi = 0$. The infection seed is *São Paulo Guarulhos International Airport*. Right: Corresponding plot in the hidden space of RWED, where the epidemic spreads as a highly correlated circular wave centered at the infection seed.

basic reproductive number $\mathcal{R}_0 = 1.5$ originated at *São Paulo Guarulhos International Airport*. In the hidden space of ED the epidemic spreads as a simple circular wave centered at the seed of the process, while in the geographical space there is no clear pattern. To validate quantitatively the goodness of the ED we use the Pearson correlation coefficient r, defined for two generic variables x and y as their covariance normalized by the respective standard deviations

$$r = \frac{\sum_i (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sqrt{\sum_i (x_i - \langle x \rangle)^2}\sqrt{\sum_i (y_i - \langle y \rangle)^2}}. \tag{4.33}$$

We then quantify the accuracy of ED measures normalized by the linearized infection speed defined by $\tau^{-1} = (\beta - \mu)$ in (2.72), with respect to the infection arrival time (4.32) for a single infection seed (*São Paulo Guarulhos International Airport*).

In Figure 4.4, each scatter point corresponds to a target airport (subpopulation) j, which is labeled infected. The high correlation with the infection arrival time found in [46], using a dominant-path approach (light-blue) is improved when considering the RWED (orange). The points on the dashed diagonal indicate a perfect agreement between the simulation and the ED. The correlation distribution considering all nodes in the network as infection seed shows that not only the measure proposed here possesses a higher averaged correlation but it is also more peaked around it, see Figure 4.5. In the inset of the same Figure we also show the very poor performance of the geographical distance.

In addition to the GMN, we perform numerical experiments on the air-traffic network of the United States of America (USA) [74] and three unweighted networks: an ER network obtained by randomly rewiring the edges of the USA dataset (with weights set all equal to one), a synthetic BA network created with $m = 5$ new edges per time step and a regular lattice (RL) with periodic boundary (i.e. a torus). The USA network includes the 500 busiest commercial airports in the United States. An edge exists between two airports if a flight was scheduled between them in 2002. The weights correspond to the number of seats available on each scheduled flight. A summary of the relevant statistical properties of the additional networks is given in Table 4.1. In Figure 4.6 (a) the comparison between the dominant-path and random-walk approach for the USA network shows that the results obtained for the GMN are robust at the country level for a single (random) seed of the infection. The correlation coefficients are $(r_{\mathrm{DP}}, r_{\mathrm{RW}}) = (0.99, 1.00)$. The overall performance with respect to the GMN is actually improved when considering all possible seeds of infection $\{i\}$, see Figure 4.6 (b) and Figure 4.5. Analogously, we find $(r_{\mathrm{DP}}, r_{\mathrm{RW}}) = (0.94, 1.00)$ when rewiring the edges of the USA network and setting the weights to one, see Figure 4.7 (a). For the other two synthetic networks, Figure 4.7 (b) and Figure 4.7 (c), we find a comparable performance between the two approaches for both BA and RL network. The correlation coefficients in this case are $(r_{\mathrm{DP}}, r_{\mathrm{RW}}) = (0.93, 0.93)$ and $(r_{\mathrm{DP}}, r_{\mathrm{RW}}) = (0.99, 1.00)$, respectively.

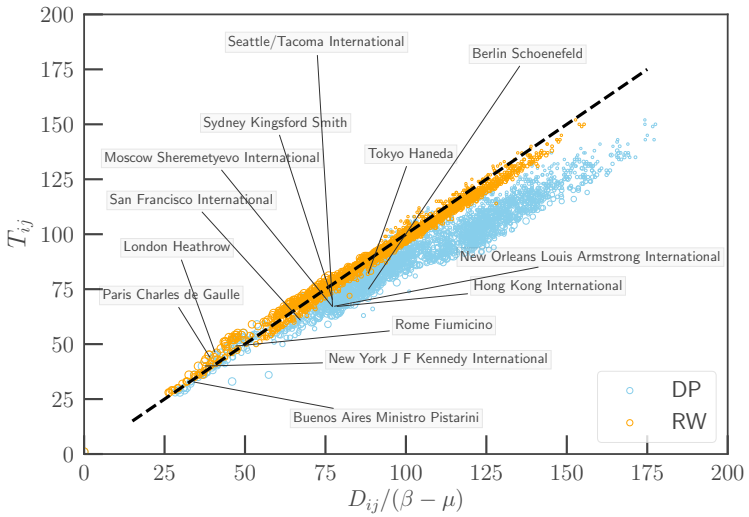


Figure 4.4: Correlation of the infection arrival times T_{ij} obtained from numerical integration of (2.89) with the dominant-path ED (light-blue) and the RWED (orange). The points on the diagonal (dashed solid line) correspond to perfect correlation. Here the infection seed i is *São Paulo Guarulhos International Airport* and each point in the scatter plot corresponds to a target airport j in the GMN, with size proportional to its strength s_j. Parameters are respectively $\alpha = 0.028$ d^{-1} (in unit of days), $\beta = 0.407$ d^{-1} and $\mu = 0.271$ d^{-1} for diffusion, transmission and recovery rates respectively. Using (4.12) this results in $\lambda \approx 1$ and a basic reproductive number of "influenza-like" diseases $\mathcal{R}_0 = 1.5$. The Pearson correlation coefficients are $r_{\mathsf{DP}} = 0.96$ and $r_{\mathsf{RW}} = 0.99$, respectively.

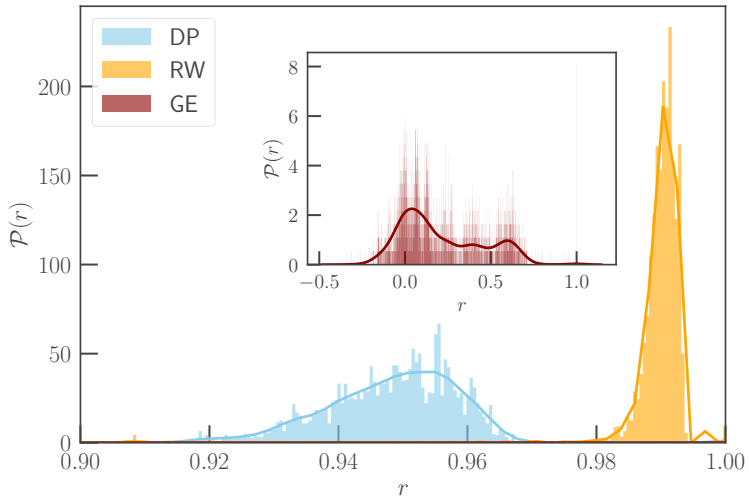


Figure 4.5: Distribution of the Pearson coefficients for all seeds $\{i\}$ and target nodes $\{j\}$ in the GMN between arrival time and ED in the dominant-path (DP) and random-walk (RW) approach. Parameters as in Figure 4.4. Inset: correlation between arrival time and geographical distance (GE).

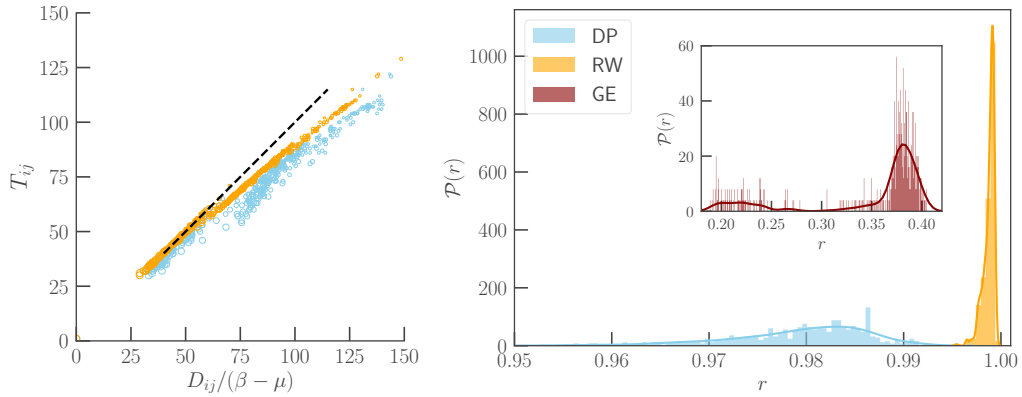


Figure 4.6: Results for the USA airport network [74]: (a) Correlation between ED (horizontal axis) using the dominant-path (light-blue) and random-walk (orange) approaches, with the infection arrival time (vertical axis). (b) Distribution of the Pearson coefficients for all seeds $\{i\}$ and target nodes $\{j\}$ in the GMN between arrival time and ED in the dominant-path (DP) and random-walk (RW) approach. Parameters as in Figure 4.4. Inset: correlation between arrival time and geographical distance (GE).

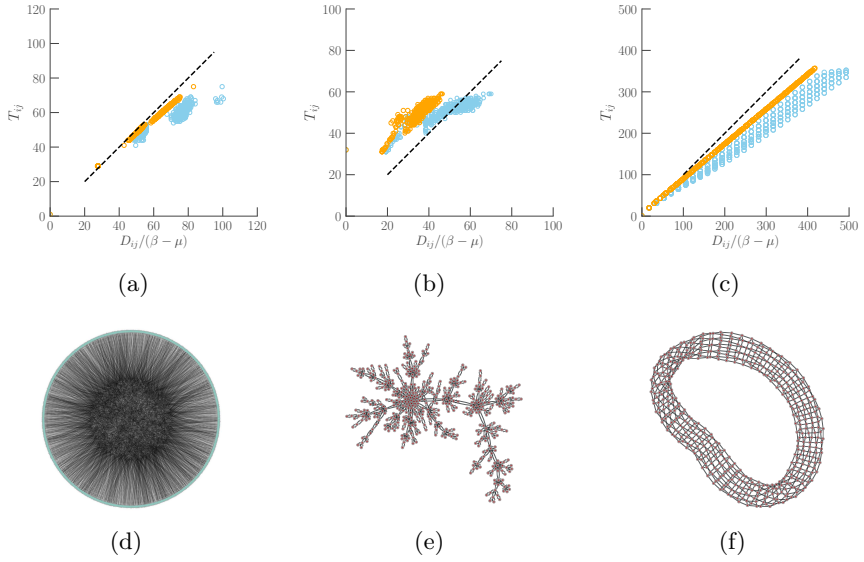


Figure 4.7: Results for the artificially constructed networks described in Table 4.1: ER (a), BA (b) and RL (c). Correlation with the infection arrival time (vertical axis) of (horizontal axis) the dominant-path ED (light-blue) and RWED (orange). Parameters as in Figure 4.4. In (d), (e) and (f) the corresponding network visualizations.

In this Chapter, we have defined the generalization of network effective distances by relaxing the assumption of simple-path propagation of spreading processes. The proposed RWED includes the previously defined dominant-path measure as a particular case. The almost perfect correlation found with the infection arrival time can be explained as follows. The contribution of looped trajectories in the propagation of information is practically all discarded because of the decreasing exponential in the walk length in (4.20). The latter damps all contributions of very long walks, and in particular allows us to neglect the contributions of infinite loops. In scenarios where multiple parallel paths are important, for instance in ER graphs or RL, the assumption of a single dominant path breaks down and the measure proposed here can be used as an efficient alternative. The predictive power of the RWED can be used for containment strategies and estimation of arrival times for real global pandemics from the underlying networks topology. The method can in fact be generally applied to any weighted and directed network besides transportation networks, e.g. social networks. It remain intriguing to test if this is the case also for rumor spreading dynamics. For unweighted locally tree-like networks both the dominant-path ED and RWED yield maximum correlation with the simulated arrival time, as the dominant path tends to dominate.

From a theoretical point of view our results show that the average infection arrival time in a metapopulation model can be quantified by the cumulant generating function of the random-walk hitting time. In fact, the generating function approach can also be used to formally derive ED from the first moment of a Gumbel distribution [120]. The connection with the cumulant generating function allows us for an interpretation within statistical physics. In particular, this would explain how the different approaches are connected in terms of the entropy associated to paths of fixed length [161, 30], see also Chapter 6. In summary, the RWED serves as a bridge between network-driven spreading processes and a generic diffusion process. Further developments and extensions include the generalization to temporal networks by considering a set of transition matrices [174].

5

Social Contagion and Leanings on Twitter

"I believe that whatever we do or live for has its causality; it is good, however, that we cannot see through to it."

–Albert Einstein

Contents

In online social platforms, users share content and can transmit news and informations to their contacts. Individuals form their opinion and make decisions influenced by their contacts in these social networks. Contrary to networks considered in the previous Chapters, many social networks, e.g. Twitter, are inherently directed and the

symmetrization of the adjacency matrix is not viable. This has profound consequences for dynamical processes and flow of information in such systems as many paths between nodes do not exist in both directions.

In the last years, on-line social media have reached a fundamental role in the political discussion as they allow to easily spread a slogan or a political campaign in a large population of users. Every time we access a piece of information, share a content or comment on a news we leave a permanent digital trace, which can be used to infer our opinion regarding one particular event or topic [77, 76, 67]. Among the variety of social media services, the micro-blogging platform Twitter is probably the most commonly used for political debate and by political leaders. This platform had a relatively recent diffusion in Italy, where it is used by nearly the 10% of the Italian adult population and where the former prime minister Matteo Renzi was the first Italian politician to substantially found his public communication on tweets (the 140 characters-long messages published on Twitter). Also, Twitter stream of data is still available to the public through its application programming interface (API), making it the best candidate to study and describe the online political debates in a country.

In fact, previous works (mostly focused on the U.S.) found that the political discussion on Twitter can be a good proxy for the overall opinion of the population, and that Twitter data can therefore be used to infer the effects of a political event or to accurately predict the outcome of a vote [42]. Moreover, several studies have highlighted the non-trivial dynamics of the opinions on Twitter and within political blogs, showing that politically active web users tend to aggregate in homogeneous communities, divided by political ideas, and avoiding discussion with the counterpart [264, 265, 76, 77].

In this Chapter, we analyze the political debate on Twitter regarding the Italian *Referendum Costituzionale* (constitutional referendum) held on December the fourth, 2016. To this end, we firstly collected vote-related tweets during the intense political discussion that took place on social media during the three months before the vote. We then analyzed these tweets employing machine learning and network theory techniques. To the best of our knowledge, we are the first to apply sentiment-analysis to the political discussion on Twitter in the Italian scenario. Additionally, we explore the possibility to use Twitter data as a reliable opinion poll and we assess their predictive power of the final outcome of the vote. We found the Italian political discussion to be no different from the more studied U.S. case: strongly polarized communities act as echo-chambers and internally speak via retweets. On the other hand, the inter-community discussion is mainly based on mentions used to report and criticize adversaries' quotes. We also found that the temporal network structure generated by such contacts, see Section 2.1.1, does react to major events happened during the political campaign. Examples include debates between the two major parties leaders held on TV and political or juridical events connected to the referendum.

The Chapter is structured as follows. We first describe in Section 5.1 the procedure adopted to retrieve the tweets and the development of the classifier used to predict the

leaning of each tweet in the dataset. From this vantage point, we define a procedure to assign a dynamical opinion to a given user in Section 5.2. These dynamical opinions are then used to characterize the temporal network of contacts. We further leverage on the reconstructed users opinion and compare our recreated opinion trend to the empirical signal given by a large set of official polls, finding a very good agreement between the two. Finally, by comparing the correlation of standard centrality measures (see Section 2.1.2) with the ability of an user to spread a rumor to other users, in Section 5.3 we identify the most influential spreaders in the online political debate. The results presented in this Chapter are discussed in [35].

5.1. The political discussion network

5.1.1. Data collection and tweets classification

Starting from the midnight of the 30th of August 2016 we collected all the tweets in Italian containing one or more of the following strings: *renzi, iovotono, iovotosi, referendum, referendumcostituzionale, bastaunsi, #NO, #SI, riformacostituzionale, m5s, pd, costituzione*. Some of these strings are the keywords of the political campaign for the two main opposite formations (*iovotono, iovotosi, bastaunsi, #NO, #SI*) or the name of the two main political parties (*m5s, pd*), while others may be linked to the political debate on the referendum. To filter out tweets written in languages other than Italian we relied on the automatic twitter language detection.

The collected tweets incorporate information on the author, the time of creation, the location (if available), the text content, the *mentions* (i.e., users cited in the text by inserting *@username*), the links and urls inserted, the hashtags, and the tweet kind. The latter specifies if the tweet is an original tweet (i.e. a new content generated by the user) or a *retweet*, meaning that the user reposted on its account a tweet previously generated by another user. A retweet contains a unique identifier of the original tweet being reposted, its original text and author. We did not reconstruct the follower network of the users (i.e., the network where two users are connected if one of the two is following the other) as we only use the single retweets/mentions as a proxy for the social interactions between individuals. Data were recorded until the midnight of December the 4th 2016, with the exception of three days due to technical reasons, right after the end of the consultation and the publication of the first exit polls. The resulting data set consists in 6 894 389 tweets, authored by a total of $N = 266\,437$ users, see Figure 5.1 for the locations of a representative sample. Note that we did not filtered for accounts possibly associated to bots, i.e. accounts not run by humans but programmed to automatically post and share information. These are known to be part of the traffic volume [231, 105, 252], but here we assume that their net effect on the contact network is negligible.

Given the political event under investigation, tweets need to be classified as belonging to one of four following classes: *irrelevant*, *pro-no*, *neutral*, and *pro-yes*. Following [240],

Figure 5.1: The locations (in red) of the collected tweets during the period starting from the midnight of the 30th of August 2016 and ending on midnight of election day (December the 4th 2016), right after the end of the consultation and the publication of the first exit polls. Each red dot corresponds to a fraction of users activity at any point in the observation time. The data shown here is a representative sample (3764 tweets) of the total collected tweets corresponding to users that actually had the Global Positioning System activated during the time when the tweet was generated.

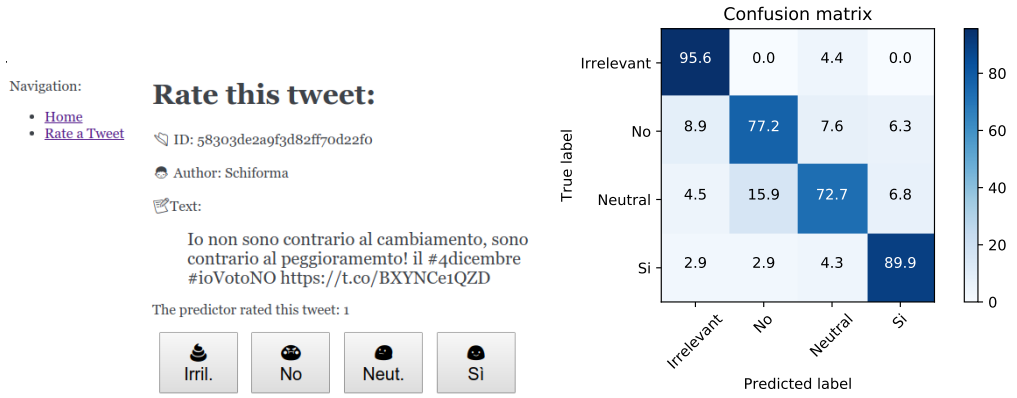
we rely on supervised machine learning techniques to classify the tweets, for which we first needed an annotated data set to train a classifier. Manual classification was made by developing a simple web interface, shown in Figure 5.2 (a), that prompts the user to classify each tweet into one of the four given categories. The information displayed to the user are: the text of the tweet, the contained *hashtags*, the author name, and all the urls included in the tweet.

Because of the sometimes ambiguous inclination of a tweet and the subjective perception of a tweet being more prone to the pro-yes or pro-no classes, we strengthened the manual classification of tweets by considering a tweet to be classified when (i) it received at least 3 votes and (ii) at least two thirds of them are in agreement. At the end of this procedure we ended up with a fairly balanced training set: out of the 1150 labeled tweets, 306 have been voted as irrelevant, 375 pro-no, 203 as neutral and the remaining 266 as pro-yes.

5.1.2. Sentiment analysis

While previous works on the Italian political scenario limited their attention to the volume of tweets containing fixed hashtags as a proxy for the overall volume of political affiliation [98, 55], we apply for the first time sentiment-analysis to Twitter data of the Italian political discussion. Although the former approach is certainly pioneering, it has several counter-effects. For example, a pro-yes tweet containing the official pro-no hashtag would straightforwardly increase the pro-no volume. Our method, detailed in the following, overcomes such difficulties.

The features of our model are the low-case words extracted from the text of the tweets where we removed punctuation and we substituted all the posted urls with their principal domain name (e.g., a link pointing to http://www.newspaper.it/politics/



(a) (b)

Figure 5.2: (a) The web interface presented to the human voter containing the unique identifier of the tweet in the database, the author's nickname and the text of the tweet. If the tweet already features a preliminary classification, this is shown above the four buttons to classify the current tweet. Once the user inputs its preference, the system automatically presents a new tweet to be categorized. (b) The confusion matrix with percentage values for the random forest model with 21 estimators using the top 200 words and hashtags as features.

`news_about_referendum` is replaced by `newspaper.it`). Smileys are left as UTF-8 characters and we leave hashtags as separate entities, i.e. we do not treat them as simple words. This last choice can be justified a posteriori by observing that the classifiers accuracy drops significantly when projecting hashtags to words by removing their leading pond #. We further process the features by stemming (using the `WordNetLemmatizer` from the nltk module of Python [37]) and removing the Italian stop-words from the corpus (using the `stopwords` collection from `nltk.corpus`). The resulting dictionary is composed by 535 different hashtags and 3773 words.

The manually annotated data set is used to train and validate three different classifiers: logistic regression, naive Bayes, and random forest [240, 2]. Results are then compared to choose the best-performing model. The performances of each model are evaluated using a 4-fold cross validation scheme: a fourth of the labeled tweets is randomly selected for validation and the remaining three fourths for training, and the procedure is repeated four times. The accuracy of a model is defined as the 4-fold average of the percentage of tweets in the validation set whose automatic classification correctly matches the manual one. The three classifiers give relatively similar performances: 80% accuracy for the random forest model, 82% accuracy for logistic regression and 79% accuracy for the naive Bayes classifier.

To gain better performances we further improved the random forest classifier by leveraging on the weights of the fitted model that naturally measure the importance of each

word. The procedure is implemented as follows:

(i) we trained a random forest classifier using the 3773 words in our dictionary (hashtags excluded) as features. The model fitted in this way has relatively low predictive power, with accuracy $\sim 50\%$;

(ii) we selected the top 200 words ranked by feature importance (the top 10 most predictive words are *referendum, renzi, no, sì, riforma, m5s, salvini, paese, grillo*);

(iii) finally, we trained a random forest model using all the hashtags and the top 200 words as features, obtaining 86% accuracy using 21 estimators and an impurity split equal to $3.1 \cdot 10^{-6}$.

The confusion matrix of the model on the annotated dataset is shown in Figure 5.2 (b). The model was finally used to classify all the remaining tweets, resulting in $\approx 2.6M$ irrelevant, $\approx 1.7M$ pro-no, $\approx 1.1M$ neutral and $\approx 0.5M$ pro-yes tweets.

5.2. Opinion dynamics

5.2.1. User dynamical opinion

In this section, we show and discuss the procedure used to define the opinion of each user in each day of the data collection, based on the users' classified tweets. We denote by $i \in \{1, \dots, N\}$ the index identifying a given user, by $T = 97$ the length of the data set expressed in days, and by $t \in \{1, \dots, T\}$ a given day ($t = 1$ corresponds to August 31st and $t = 97$ to December the 4th, the date of the referendum). Then, we denote by $\mathcal{N}_i^{\mathrm{Y}}(t)$ the number of pro-yes tweets posted at time t by author i, and $\mathcal{N}_i^{\mathrm{N}}(t)$ analogously for the pro-no tweets. Note that tweets classified as irrelevant are considered equivalent to those classified as neutral in what follows.

We define the daily activity as the opinion that the user expresses the most in its daily tweets, namely

$$a_i(t) = \begin{cases} +1 & \text{if } \mathcal{N}_i^{\mathrm{Y}}(t) > \mathcal{N}_i^{\mathrm{N}}(t) \\ 0 & \text{if } \mathcal{N}_i^{\mathrm{Y}}(t) = \mathcal{N}_i^{\mathrm{N}}(t) \\ -1 & \text{if } \mathcal{N}_i^{\mathrm{Y}}(t) < \mathcal{N}_i^{\mathrm{N}}(t) \end{cases} \tag{5.1}$$

Note that this choice also implies that if user i either posts only neutral/irrelevant tweets or does not tweet at all, then its daily activity is assumed to be neutral.

The daily activity defined in (5.1) is not suitable to represent a user's opinion, because it only captures what the user tweeted on a given day. Indeed, it is reasonable to assume that a person maintains her/his opinion even though she/he does not declare it on Twitter every single day. For such a reason, we define an opinion function that maps the daily activity of user i to $o_i(t)$, the actual opinion of user i at time t. To include a

memory effect, the opinion of user i is the projected weighted sum of its daily activity up to time t

$$o_i(t) = \mathbb{P}\left(\sum_{s=1}^{t} w(s,t) a_i(s) \right), \tag{5.2}$$

where the projection function $\mathbb{P}$ is discussed below and the $w(s,t)$ are exponentially decaying weights given by

$$w(s,t) = \frac{e^{-(t-s)/\tau}}{\sum_{s=1}^{t} e^{-(t-s)/\tau}}. \tag{5.3}$$

Here, the characteristic time scale τ gives an approximation of the typical number of days before time t that are taken into account when determining the opinion of the user. The intuition behind (5.3) is that the more recent a tweet is, the more it affects the opinion value of the user.

The function $\mathbb{P}$ projects the argument within the parenthesis in (5.2) onto the discrete opinion set $\{-1, 0, +1\}$. A natural choice for such projection is therefore the sign function. However, an undesirable drawback of setting $\mathbb{P}(x) = \text{sign}(x)$ is that even a mildly polarized user who tweeted in favor of, say, the yes vote on the very first day and then always posted neutral tweets for the following three months, would maintain a $+1$ opinion forever. Indeed, even though weights $w(0,t)$ rapidly approach zero as t becomes large, a sign function would still project the resulting small (but positive) sum on a pro-yes opinion. Then, we consider the projection to be a sign-like function with a ϵ-widened preimage of zero, i.e.

$$\mathbb{P}_\epsilon(x) = \begin{cases} +1 & \text{if } x \in [\epsilon, 1] \\ 0 & \text{if } x \in (-\epsilon, \epsilon) \\ -1 & \text{if } x \in [-1, \epsilon] \end{cases} \tag{5.4}$$

Overall, two parameters are involved in the definition of the users opinion (5.2): the decaying time of the weights τ and the widening parameter of the projection function ϵ. Since we set the parameter τ to be the typical number of days that a user retains its opinion, we infer it from the data as follows. We fix τ to be the average number of days between two coherent tweets authored by the same user which are not separated by a differently classified tweet. Considering that on average users tweet coherently every 5 days and 7 hours (without showing any different daily activity in between), we deduce that the personal opinion is preserved for $\tau = 5.29$ days on average. On the other hand, the role of the parameter ϵ is less self-evident. We recall that such parameter is introduced to allow users to return on a neutral position after a period of inactivity, but it actually has a more complex effect on the time evolution of the users opinion $o_i(t)$ under several aspects. We first observe that ϵ also has an effect on the memory of

users' past tweets. To choose a proper value of ϵ, we tested the outcome of the global process considering different values of ϵ in the interval $[0, 0.1]$. First, we consider the percentage of users that forever maintain a polarized opinion after they assume one, who we name *stubborn users*. We found that the effect of increasing the parameter ϵ is, on the one hand, to reduce the users' stubbornness (the tendency to maintain a given opinion once it is assumed). On the other hand, a larger value of ϵ results in a smaller number of abrupt jumps between Yes and No, at the cost of larger opinion variability. However, we still lack a quantitative method to asses the value of ϵ that produces a good trade-off between users' stubbornness and opinion volatility. Indeed, since stubbornness and opinion retention are related, smaller values of ϵ result in longer users' memory. In order to quantify such memory length, we considered the mean and the mode of the number of days an opinion is preserved once it is assumed. Such statistics are computed solely considering those users that acquired an opinion only once in the dataset, and then lost it. We not consider users that assume and lose an opinion multiple times in order to focus on the effect of small clusters of Tweets which are frequent enough not to let the opinion to vanish. Also, the requirement that the opinion has to be abandoned eventually rules out users that are ideologically polarized and would significantly extend the memory length statistics. We found that only for $0.07 < \epsilon < 0.08$ the memory length is reasonably close to the chosen value of $\tau = 5.29$ days defined above. Hence, we analyzed the opinion dynamics of some users who only tweeted sporadically in favor of either Yes or No, and such analysis corroborated the choice of $\epsilon^* = 0.075$. As a synthetic description of the opinion resulting from the choice of $\tau = 5.29$ days and $\epsilon = 0.075$, we find that the empirical distribution of the users aggregated opinion

$$\overline{o}_i = \frac{1}{T} \sum_{t=1}^{T} o_i(t), \tag{5.5}$$

exhibits an evident skewness towards negative values (i.e. towards pro-no opinions). Motivated from the previous discussion we set $\epsilon = 0.075$.

To illustrate the resulting opinion time course, we plotted the daily activity histograms of a few users and the resulting opinion time course in Figure 5.3. Here the y-axis represents the time spanning from $t = 1$ (31st August 2016) to $t = 97$ (the date of the referendum), whereas the z-axis counts the number of tweets authored by the user on a specific day. Such histograms are subdivided according to how the tweets were classified. In fact, the x-axis contains the possible tweets classifications: -1 for a tweet supporting No (in red), 0 for a neutral or irrelevant tweet (gray), and $+1$ for a tweet supporting Yes (blue). Figure 5.3 (a) shows the histogram for the daily activity of the official pro-yes committee's account @bastaunsi (twitter id 733695386846662657), whereas Figure 5.3 (b) represents one of the official pro-no committees' account @comitatono (twitter id 696674734969397248). The line lying on the right side of the xy-plane shows the resulting user's opinion time course, where blue stays for a pro-yes opinion, gray for a neutral

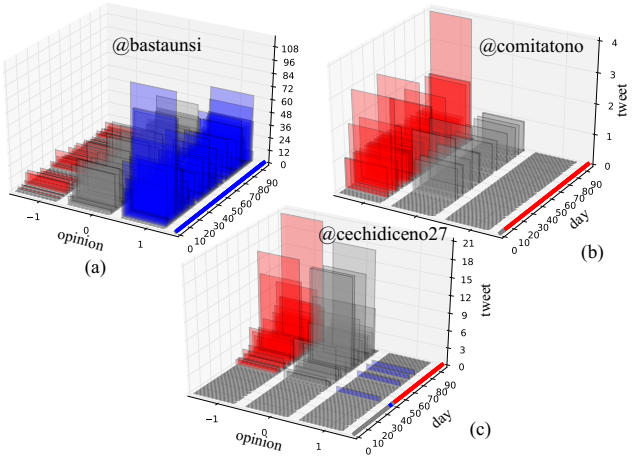


Figure 5.3: Histogram of the number of tweets authored by different users, subdivided in number of tweets classified as pro-no (-1, red), neutral or irrelevant (0, gray), and pro-yes ($+1$, blue). The colored line on the right of each panel shows the resulting opinion of the user as defined in (5.2) with $\tau = 5.29$ days and $\epsilon = 0.075$. Panel (a) refers to the official pro-yes committee's account (@bastaunsi), panel (b) to the official pro-no committee's account (@comitatono), and panel (c) to user @cechidiceno27 that exhibits an opinion switch from a temporary pro-yes to a sustained pro-no leaning.

opinion, and red for a pro-no opinion. For both accounts we can observe a continuous twitting activity supporting the relative pole. As expected, such a sustained activity results in an opinion which is neutral at the beginning (neither of the accounts twitted on August 31st) but it changes for good once the first polarized tweet is posted. It is interesting to remark that the account @bastaunsi produced a much larger number of tweets than @comitatono, with a maximum of over one hundred tweets per days for the official pro-yes account versus a maximum of four tweets per day for the official pro-no account. Moreover note that, although the machine-learning algorithm classified some of @bastaunsi tweets as pro-no (note the red bars on the left of Figure 5.3 (a), the large mole of pro-yes tweets allows to overlook what is likely to be a misclassification of the machine-learning algorithm. Finally, Figure 5.3 (c) shows the daily activity histograms for account @cechidiceno27 (twitter id 780796377148162048). We observe that such user only started posting at $t = 40$ (9th October 2016) with a pro-yes tweet, which resulted in a pro-yes opinion that lasted three days. Then, because of the repeated pro-no activity, its opinion returns to be neutral for a couple of days, just to settle definitively on a -1 opinion at $t = 45$.

Now that we defined the procedure to evaluate the opinion of each user at a given time we can investigate how the topology of the network of contacts among Twitter users is shaped by the leaning of the users themselves.

5.2.2. Comparison with official polls

Starting from our dataset of annotated tweets we build two different temporal networks: the *retweet network* (RN) and the *mention network* (MN). For the RN we assign a directed edge $j \to i$ when user i retweets user j. Contrary for the MN we assign

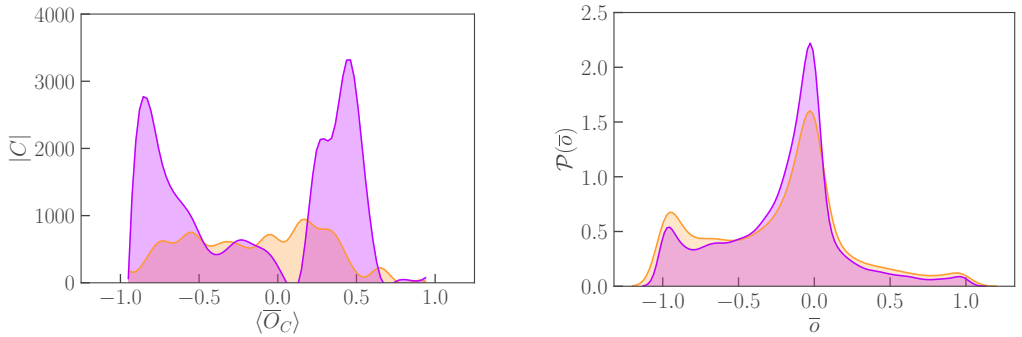
a directed edge $i \to j$ when i mentions j in a tweet. The two conventions on the edge direction are adopted so as to reproduce in the synthetic system the information flow found in the empirical social network. The topology of these networks reflects the structure of the political debate within Twitter. In the following we first detect the communities present in the system aiming at the analysis of their political polarization and investigate how the users' political opinion is distributed within them. Finally, we characterize the preferred mean of communication between communities with diverse average opinion.

In order to gain further understanding of the political structure of the networks we implemented a community detection algorithm, applying the Louvain method [39]. The Louvain algorithm is a greedy optimization method that finds the optimal division of a network into communities by iterative optimization of the network modularity, a measure of the density of edges between nodes within a community with respect to the density of edges outside that community. The method returns a set of subgraphs more densely connected to one another than to other nodes, along with a hierarchy of communities at different scales. Studying the size of the detected communities with respect to the average opinion inside the community gives an insight on the network political composition. We measure the average opinion $\langle \overline{O}_C \rangle$ inside a community C as the mean value of the opinion $o_i(t)$ among the users belonging to the C community and over the full observation period. Thus, the community average $\langle \overline{O}_C \rangle$ is obtained by averaging (5.5) over the set of nodes belonging to community C

$$\langle \overline{O}_C \rangle = \frac{1}{|C|} \sum_{i \in C} \overline{o}_i, \tag{5.6}$$

where $|C|$ is the number of nodes in community C. The result of such analysis is shown in Figure 5.4 (a). In the RN the larger communities are those with a strongly polarized average opinion, whereas we find the communities in the MN to have a weaker opinion polarization (the larger MN communities feature $\langle \overline{O}_C \rangle \approx 0$). Interestingly, we find the opinion polarization to be stronger for the pro-no communities of the RN, as can be seen by comparing the relative shift with respect to zero average opinion of the two violet peaks in Figure 5.4 (a). These communities are generally larger and more polarized (i.e. with more negative opinion) with respect to the ones supporting the pro-yes faction. Moreover, we observe also in the MN case a slight shift of the communities opinion toward the negative pole. This shift is reasonably due to the overall predominance of the pro-no side which were against the governing party (which was, on the contrary, promoting the pro-yes side), as can be seen in the distribution of the users time-averaged opinion shown in Figure 5.4 (b). A visual representation of the SCGC for the mention aggregate (time-averaged) network with each user average opinion (5.5) is shown in Figure 5.5.

Next, we compare the daily temporal evolution of the average opinion obtained from the Twitter dataset using the sentiment analysis described in the previous Section and the opinion trend obtained through the official polls. The sources considered compre-



(a) (b)

Figure 5.4: (a) The size of connected communities plotted as a function of the average opinion $\langle \overline{O}_C \rangle$ of users belonging to community C, for MN (orange) and RN (violet). In both cases, we show the 95th percentile of the community size distribution found for communities with a given average opinion. (b) Kernel density estimation of the users time-averaged opinion in the SCGC of MN (orange) and RN (violet).

hend several sets of official polls. These polls are carried out by different statistical research institutes and commissioned by different customers such as: the official website of the political and electoral polls of the Italian Government, popular Italian newspapers such as *La Stampa*, *Il Corriere della Sera*, and *Il Sole 24 ore*, and private monitoring companies. The sample size of the surveys is heterogeneous but always statistically significant, from a minimum of 400 to a maximum of 4000 individuals (about 1100 people interviewed on average). In addition, the time span of the official polls we considered goes from the 31th of August (the day we started collecting the data) to the 17th of November 2016, the day before the pre-election silence started. This is because the Italian law forces the official polls to stop two weeks before the vote. On the other hand, the Twitter data were recorded until the midnight of the 4th of December.

In our analysis, we compare the daily average opinion $\langle o(t) \rangle = N^{-1} \sum_{i=1}^{N} o_i(t)$ with the official polls opinion, defined as the difference between the proportion of pro-yes and pro-no voters in each poll. In both the Twitter opinion and the official polls we considered also the undecided users. Moreover, while the average opinion $\langle o(t) \rangle$ measured from the tweets has a daily temporal resolution (as we project the daily user activity), the official polls do not have a regular frequency as they were generally published every couple of days.

In Figure 5.6 we show the qualitative comparison, at the maximum temporal resolution, between $\langle o(t) \rangle$ and the opinion reported by the official polls. In this plot, zero represent the perfect equilibrium between the pro-yes and pro-no voters and a positive or negative value of the average opinion represent respectively a majority of pro-yes or

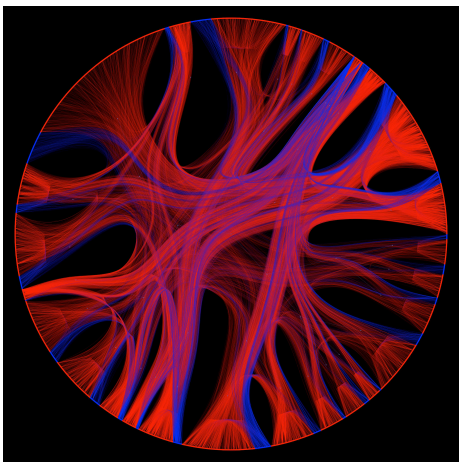


Figure 5.5: Circular visualization of the SCGC aggregated MN. The time-averaged opinion $\overline{o}_i$ of each user is represented with color codes as blue (pro-yes) if $\overline{o}_i > 0$ and red (pro-no) if $\overline{o}_i < 0$, and analogously for the edges. The node's ordering is given by the stochastic block model [221]. A pattern of segregation between local pro-yes and pro-no communities is clearly visible while the overall exchange in links between opposite political opinions is very low, confirming the high level of segregation found for the community average opinion $\langle \overline{O}_C \rangle$ in Figure 5.4.

pro-no voters. We observe that, during the first sampling period from August 31th to September 20th, the average opinion obtained through the official polls fluctuates around zero, reaching a maximum of 0.09 when the major of Rome, who endorsed the No, was involved in legal issues (red bar in Figure 5.6). On the other hand, the average opinion obtained through Twitter data starts near zero and then fundamentally decreases toward negative values. After the 20th of September, when the Italian government fixed the official voting day, the behavior of the official polls starts to be more stable and prone to a pro-no vote. On the 5th of October, when the regional administrative court (*tribunale amministrativo regionale*, TAR) of region Lazio rejected a petition which had requested a partial invalidation of the referendum, the trend changes in both cases towards a more pro-yes political orientation. Subsequently, the two opinions approach more negative values until the 17th of November, when the official polls opinion is equal to $-0.07(\pm 0.03)$ while $\langle o(t) \rangle = -0.13$. To compare the different trends in a quantitative way, we further re-sampled the two time series averaging them with a weekly timescale. The Pearson coefficient (4.33) between the two time series is $r = 0.888$ with a p-value equal to $p = 10^{-4}$, showing a high degree of correlation between the two time series. Moreover, we find our latest reconstructed opinion value ($\langle o(t) \rangle = -0.15$) to outperform the official polls in predicting the final referendum result. Indeed, given that 65.47% of the eligible population voted and 40.88% of this percentage voted Yes and the remaining 59.12% voted No, the average opinion on the vote outcome is -0.12.

5.3. Rumor spreading

After the analysis of the networks topology, it is worth considering how the edges arrangement and their activation patterns shape the evolution of rumor spreading in both MN and RN. Since in directed contact networks each node (user) can receive and spread

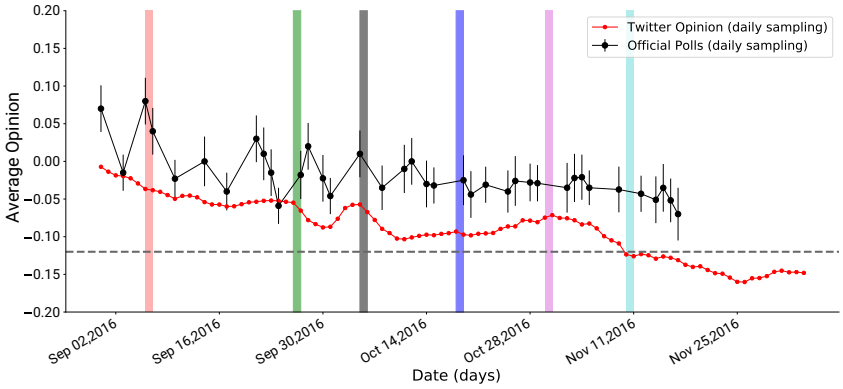


Figure 5.6: The daily comparison between the variable $\langle o(t) \rangle$ (red) and the opinion obtained by official polls (black). The error bars on the official polls data represents the statistical error range given in each poll. The black dashed line represents the final result of the voting day -0.12. The vertical bands represent some events that had a significant impact for the referendum debate: (red) the mayor of Rome, who previously endorsed the No, is involved in legal issues; (green) the Italian government fixes the referendum day; (black) the regional court of the region Lazio receives an appeal to invalidate part of the Referendum question formulation; (purple) the public debate about the referendum reaches the first pages of the main Italian newspapers; (pink) television debate with the Italian prime minister; (cyan) an important national meeting, Leopolda, organized by the Government party, is held in Florence.

a rumor only within the SCGC, we restrict our analysis to that subset only.

5.3.1. Causality of the temporal networks

Both MN and RN are temporal networks defined by an ordered set of (unweighted) adjacency matrices $\{\mathbf{A}^{(t)}\}$ with $(t = 1, \ldots, T)$. A preliminary characterization of rumor spreading is obtained by considering the aggregate representation. Here we consider the unweighted aggregate instead of the average (2.3). Then we add the snapshots using the boolean sum $\bigvee$ (operator OR)

$$\overline{\mathbf{A}} = \bigvee_{t=1}^{T} \mathbf{A}^{(t)}. \tag{5.7}$$

In Table 5.1 we report some properties for the aggregated networks of the full and SCGC MN and RN. The high value of the second moment of the out-degree distribution, observed for both MN and RN, reflects the rapid increase of the topological fluctuations in the network of interactions between Twitter users. This has important consequences for epidemic processes unfolding on such networks [215] and we expect it to play a major role also for rumor spreading.

The temporal feature of the empirical dataset presents many challenges, first and

	N	E	D	$\langle C \rangle$	$\langle k^{out} \rangle$	$\langle (k^{out})^2 \rangle$
MN	80030	643019	∞	0.22	8.03	794.17
MN SCGC	15294	367641	6	0.34	24.04	2999.35
RN	179680	1582288	∞	0.11	8.81	10771.41
RN SCGC	27437	956101	7	0.28	34.85	24074.57

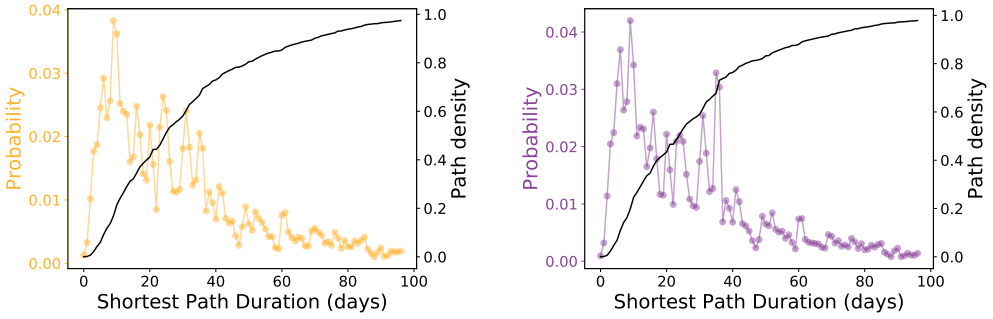
Table 5.1: Statistical properties of the full and SCGC time-aggregated MN and RN. The various quantities are: the number of nodes N, the number of edges E, the diameter D and the global clustering $\langle C \rangle$ computed from the corresponding undirected graphs, the first moment $\langle k^{out} \rangle$ and the second moment $\langle (k^{out})^2 \rangle$ of the out-degree distribution.

foremost, *causality*. To spread information from node j to node k at time $t' = t + \Delta t$, the same information must have already reached node j from node i at a previous time $t < t'$. Therefore the time aggregation of the temporal networks can include paths of rumor propagation that are not present in the causally-ordered temporal sequence of contacts.

To quantify the impact of causality, we compute the *causal fidelity* [174] $c \in [0, 1]$ of the SCGC. The latter is defined as the fraction of the number of paths in the time-aggregated static network which can be also taken in the temporal one. Thus $c = \rho(\mathcal{A})/\rho(\overline{\mathbf{A}})$, where

$$\mathcal{A} = \bigwedge_{t=1}^{T} (\mathbf{I} + \mathbf{A}^{(t)}) \tag{5.8}$$

is the *accessibility matrix* of the temporal network obtained with the boolean product $\bigwedge$ (operator AND), while $\rho(\overline{\mathbf{A}}) = \sum_{ij} \overline{A}_{ij}/N^2$ is the density of matrix $\overline{\mathbf{A}}$, and analogously for $\rho(\mathcal{A})$. The maximal causal fidelity $c = 1$ implies that the temporal and static representations share the same path density. On the contrary, a low value of c indicates that most of the paths in the static aggregate approximation do not follow a causal sequence of edges and thus do not belong to the temporal network. We find $c = 0.973$ and $c = 0.979$ for the MN and RN, respectively. These values suggest that the temporal causality-driven effects in both SCGC networks are negligible. Thus, it is reasonable to characterize the rumor dynamics only considering the time-aggregated representations of the SCGC. In Figure 5.7 (a) and (b) we show (left axis) the probability $\rho(\mathbf{A}^{(t)}) - \rho(\mathbf{A}^{(t-1)})$ to find a path of length t between two randomly chosen nodes, and the corresponding density of the accessibility graph. In spite of the networks directness, the density of the accessibility graph (black line) shows that after only half of the observation period (about 50 days) more than 80% of the network is causally connected. Therefore, the aggregated network representations give a good approximation of the temporal one. The shortest path length is broadly distributed for both the MN and RN, while its most frequent value is in both cases attained at a nine-days long path. This means that the typical spreading time scales are of the order of 10 days.



(a) (b)

Figure 5.7: Distribution of the shortest path duration (color) and the density $\rho(\mathcal{A})$ of the accessibility matrix (black) for (a) the SCGC of the MN (orange) and (b) the SCGC of the RN (violet). Causal fidelity values are $c = 0.973$ and $c = 0.979$ for the MN and the RN, respectively.

5.3.2. Spreading dynamics

Models for rumor spreading can be formulated as variants of the SIR model for disease epidemics, see Section 2.3.2, in which the recovery process does not occur spontaneously, but rather is a consequence of interactions. The basic idea behind this modification is that it is worth propagating a rumor as long as it is novel for the recipient. If the spreader finds that the recipient already knows the rumor, he or she might lose interest in spreading it any further. The formalization of this process is due to Daley and Kendall [84]. Individuals can be in one of three possible states: ignorant (S, equivalent to susceptible in SIR), spreader (I, equivalent to infected), and stifler (R, equivalent to recovered). In a slightly distinct version, introduced by Maki and Thompson [185] when a spreader contacts another agent and finds it in state I, only the former turns into a stifler, the latter remaining unchanged. The possible events and the corresponding rates are

$$\begin{aligned} S + I &\xrightarrow{\beta} I + I \\ R + I &\xrightarrow{\mu} R + R \\ I + I &\xrightarrow{\mu} R + I \end{aligned} \tag{5.9}$$

where β and μ are the spreading (transmission) and the stifler (recovery) rates. This defines the ignorant-spreader-stifler (ISS) model. As for the SIR model, starting from a single informed individual, the rumor propagates in a population of size $N = S(t) + I(t) + R(t)$ with an increase in the number of spreaders. Asymptotically, all spreaders turn into stiflers and in the final state there are only ignorants or stiflers. The order parameter of the model is the *reliability*, defined as the fraction of stiflers (removed for

SIR) in this asymptotic state

$$\rho^R(\infty) = \lim_{t\to\infty} \frac{R(t)}{N}. \tag{5.10}$$

The mean-field solution where individuals are assumed to be well mixed and to interact with each other completely at random is given by the transcendental equation [23]

$$\rho^R(\infty) = 1 - \exp\left(-\left(1 + \frac{\beta}{\mu}\right)\rho^R(\infty)\right). \tag{5.11}$$

It follows that, $\rho^R(\infty)$ is positive for any $\beta/\mu > 0$, i.e. the rumor spreads macroscopically for *any* value of the spreading parameters, at odds with what happens for the SIR dynamics, which has the mean-field finite threshold $\widetilde{\beta}_c = 1$, where $\widetilde{\beta} = \beta/\mu$.

In general, models for epidemic spreading are strongly affected by very heterogeneous topologies, so it is natural to ask what happens for rumor dynamics. When the Maki-Thompson model is simulated on scale-free networks it turns out that heterogeneity hinders the propagation dynamics by reducing the final reliability, still without introducing a finite threshold. Why this happens is easily understood: large hubs are rapidly reached by the rumor, but then they easily turn into stiflers, thus preventing the further spreading of the rumor to their many other neighbors. If spontaneous recovery is also allowed, justified as the effect of forgetting, it turns out that the model behaves exactly as SIR: macroscopic spreading occurs only above a threshold inversely proportional to the second moment $\langle k^2 \rangle$, see (2.78).

The directness feature of both MN and RN plays a fundamental role in the corresponding evolution of the ISS dynamics as many users will never be able to spread a rumor that can reach a substantial portion of the network. In the Twitter case, spreaders correspond to users that have an information (such as specific news about the referendum) and if one spreader meets an ignorant user, then the latter begins to spread this rumor as well. On the other hand, stiflers are users that lose interest in the news and persuade spreaders to stop propagating the rumor. Finally, the presence of two spreaders together could bring one of them to became a stifler.

5.3.3. Influential spreaders on Twitter

Understanding the impact of individual nodes on the function of a network is one of the most fundamental and open problem in network science. Several studies have addressed the question, starting with the seminal work of Kitsak *et al.* [159], where the authors showed how the k-shell decomposition of the graph, see Section 2.1.2, can capture how central a node is with respect to its ability to spread information to the rest of the network. The lack of a satisfactory understanding of the problem comes from the high heterogeneity of the role played by the individual nodes in a complex network. Centrality

measures are then used to quantitatively derive the importance of individual nodes [272].

Here we apply standard heuristic centrality measures: the out-degree k^{out}, betweenness c^B [109], closeness c^C [237], eigenvector centrality c^E [40] and k-core index k_c [246], defined in Section 2.1.2 and PageRank centrality x [210], defined in Section 2.2. Other non-heuristics centrality measures (as the non-backtracking centrality [192], which is supposed to match exactly the spreading ability at criticality for the SIR model [228]), cannot be used on directed networks. Because of the highly directed nature of both RN and MN we do not consider such metrics here.

The high causal fidelity values found for the SCGC networks suggests to consider the topology as frozen during the time evolution of the system. The static assumption allows us to use standard static centrality measures to rank the users' influence. Given a set of initial rumor spreaders (seeds), we are interested in quantifying the dependence of those seeds' out-degree $k_i^{out} = \sum_j A_{ij}$ on the final outbreak size, varying β and μ. The natural measure of the nodes outbreak size is the *spreading ability* q_i of node i, defined as the average number of nodes in the stifler state (recovered) at the end of the spreading process. Thus, the spreading ability is the average reliability defined in (5.10)

$$q_i \equiv \langle \rho^R(\infty) \rangle_i \,, \tag{5.12}$$

where the average $\langle \ldots \rangle_i$ is evaluated over 10^2 different realizations of the stochastic ISS dynamics described above, with the rumor originated by the seed user i.

Initially all users are ignorant, except the seed i. At each time step, spreaders spread the rumor to their ignorant neighbors with probability β and turn into a stifler with probability μ for each stifler neighbor. The dynamics stops when there are no more spreaders. We choose $\beta = 0.1$ and $\mu = 1.0$ as the reference observation point in the parameter space, because this is the closest at the decimal precision to the epidemic threshold, assuming that it is vanishing, see (5.11). The top-10 users, ranked in terms of their spreading ability (5.12), are given in Table 5.2 for the MN and RN, respectively. Surprisingly, the top spreaders are only low-profile users and no account of important news agency or political figure is present.

In Figure 5.8 we show the kernel density estimation of the spreading ability with the relevant network centralities. We find that the best performing measure is the out-degree k^{out}, followed by the closeness centrality c^C and the k-core index k_c for both MN and RN. To compare the rankings obtained by the users spreading ability in the dynamics and the centrality measures we use the Spearman's rank correlation coefficient r, defined for two generic rankings as their covariance normalized by the respective standard deviations, i.e. as the Pearson' correlation (4.33) for the respective rankings. The correlation coefficient in the full spreading-probability range with immediate recovery ($\mu = 1$), is given in Table 5.3. Besides the very bad performance of PageRank for both MN and RN, the overall best-performing metric to rank users is the out-degree. Our results show also that all metrics besides PageRank perform better in the MN with respect to the RN. For

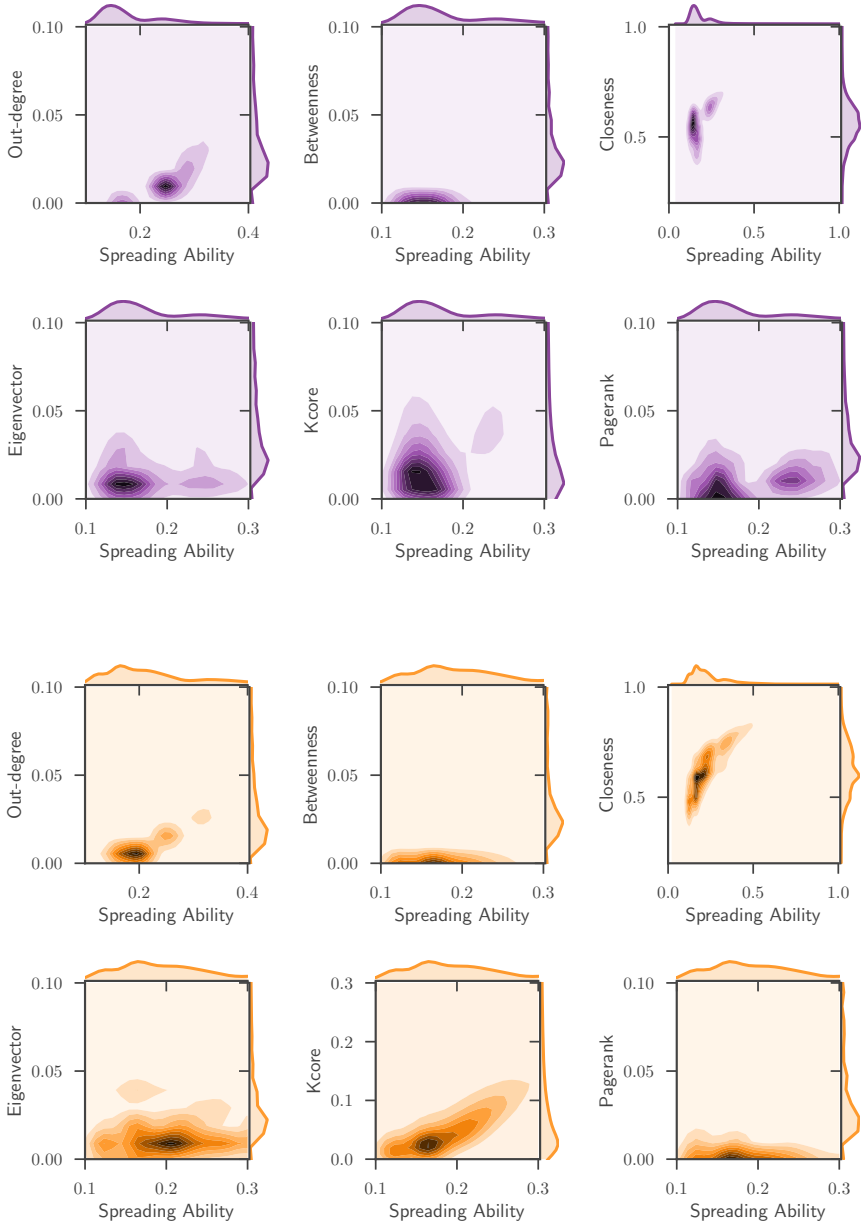


Figure 5.8: Kernel density estimation of the correlation between the distributions of the max-normalized spreading ability $q/\max[q]$ and the max-normalized centralities for the aggregated SCGC MN (orange) and RN (violet). Parameters are $\beta = 0.1$ and $\mu = 1.0$.

Rank	UserID (MN)	UserID (RN)
1	@DartSirius	@Dani_Gambit
2	@guffanti_marco	@nuccioaltieri
3	@lorenzo3107	@cadolo56
4	@Alessandro02088	@attanasio_g
5	@mdpennalunga	@GiuliaPozzuoli
6	@alessandrab72	@nonfraledonne
7	@onda_di_mare	@Bessico2
8	@angeloargento	@rpp_tweet
9	@fcerasani	@LaVarcaDiNoe
10	@pasqualegranata	@DCrognaletti

Table 5.2: Top-10 spreaders for the MN (left) and the RN (right), ranked with their spreading ability (5.12) for transmission and recovery rates $\beta = 0.1$ and $\mu = 1.0$, respectively.

the RN all metrics performed very poorly, which could be due to the strong polarization of the network compared to the MN. For the MN a local quantity such as the out-degree is the best measure to rank users as it displays an almost perfect correlation with the ranking of the spreading ability.

In this Chapter, we measured and characterized the discussion about a political event of national relevance in Italy using data from the Twitter microblogging platform. We discussed the procedure implemented to collect tweets related to the Italian constitutional referendum, which allowed us to obtain a large amount of data for our analysis (approximately 7 millions tweets). Using a manually annotated subset of tweets, we trained a classifier able to predict the leaning of tweets with great accuracy (86% accuracy in a 4-fold cross validation). We deployed this classifier to predict the opinion of each user in the system given the user's history and activity. Notably, our definition is dynamical so that the opinion of an user can change in time and it is not bound to a value computed at the end of the observation period. Thanks to the dynamical opinion, we performed a characterization of the interaction network topology in terms of the average opinion. We found strongly polarized communities composed by users sharing the same opinion that internally interact with retweets, and that interact with other communities only by mentions. Regarding the opinion trend, we found our estimate to be in good agreement with official polls. Our method is particularly interesting as it is significantly cheaper to track twitter activity rather than to finance a poll. Moreover, Italian laws prohibit companies and parties to perform and publish polls in the two weeks preceding a vote whereas, to the best of our knowledge, no restriction is currently given on Twitter data. It is therefore possible to track and characterize the general opinion dynamics in Twitter up to the date of an election, and for a longer time span with respect to the official surveys. Besides the opinion trend, our analysis also allows for the identification

	Mentions Network (MN)						Retweets Network (RN)					
β	k^{out}	c^B	c^C	c^E	k_c	x	k^{out}	c^B	c^C	c^E	k_c	x
0.1	**0.98**	0.63	0.83	0.36	0.83	0.15	**0.49**	0.38	0.48	0.20	0.45	0.22
0.2	**0.97**	0.63	0.83	0.35	0.83	0.15	**0.48**	0.37	0.46	0.20	0.44	0.22
0.3	**0.97**	0.63	0.83	0.35	0.82	0.15	**0.48**	0.37	0.46	0.19	0.44	0.21
0.4	**0.96**	0.62	0.82	0.35	0.82	0.15	**0.48**	0.37	0.46	0.19	0.44	0.22
0.5	**0.96**	0.62	0.82	0.36	0.82	0.15	**0.48**	0.37	0.46	0.19	0.44	0.21
0.6	**0.95**	0.62	0.81	0.35	0.81	0.15	**0.48**	0.36	0.46	0.18	0.43	0.21
0.7	**0.94**	0.61	0.81	0.35	0.80	0.15	**0.49**	0.37	0.46	0.20	0.44	0.22
0.8	**0.93**	0.61	0.80	0.35	0.80	0.15	**0.49**	0.37	0.46	0.19	0.43	0.22
0.9	**0.92**	0.60	0.79	0.34	0.79	0.15	**0.48**	0.36	0.45	0.19	0.43	0.21
1.0	**0.92**	0.60	0.79	0.34	0.78	0.14	**0.48**	0.36	0.45	0.19	0.43	0.21

Table 5.3: Values of the Spearman's rank correlation coefficient for the MN (left) and for the RN (right) in the full β range at $\mu = 1.0$ with the spreading ability of the out-degree k^{out}, betweenness c^B, closeness c^C, eigenvector c^E, k-core index k_c and PageRank centrality x.

of key-events influencing the overall opinion of the system. This can possibly provide for a real-time investigation of the response of the population to some public declaration or political events.

The results are twofold: on the one hand they allow us to investigate the influential spreaders and the relevant nodes of the retweet and mention networks, while on the other hand they allow for a prediction of the population opinion trend. The former point is achieved using the temporal network of interactions, thus without collecting the static network of friendships, an operation that turns out to be unfeasible on such large networks. It is worth noting that the influential spreaders identified by our method are private users and not the official pro-yes and pro-no accounts as one would expect, see Table 5.2. From the ranking correlation of the rumor-spreading ability of the active users, our analysis shows a clear out-performance of the out-degree over all other measures, in particular in the mention network where the correlation is almost perfect. Although we did not perform extensive study of other measures, among the various centralities we also find the k-core and closeness centrality as relevant for the identification of influential spreaders in social networks. Differently from previous findings [85], in the social network of political discussion analyzed in this work, the simplest local measure of connectivity in the network, the out-degree k^{out}, is sufficient to estimate the correct ranking of users with extreme accuracy, with correlation up to the value $r = 0.98$ when approaching criticality from above.

6

A New Metric for Influencers Identification in Complex Networks

> "*Per quanti sforzi potete fare per prevedere il futuro, esso vi sorprenderá.*"
>
> –Giorgio Parisi

Contents

IDENTIFYING those nodes who, once they initiate a spreading process, maximize the infected fraction of nodes, is a fundamental and unsolved problem in network science. In a seminal work [159], the authors showed that the nodes with the largest degree ("hubs" in the network science literature [17]) are not necessarily the most influential

spreaders, and nodes with fewer connections but located in strategic network positions can initiate larger spreading processes. Following that work, several centrality measures [182, 179], originally aimed at quantifying individuals' influence and prestige in social networks [153], have been tested with respect to their ability to identify the influential spreaders [63, 41, 281, 180, 85, 29, 183, 228, 213, 173]. The results of this massive effort have been often contradictory. The k-core centrality (based on the k-shell decomposition of a graph [247, 94]) has been found to outperform other metrics in [159] and subsequently in [219, 85]. This conclusion has been challenged in a number of studies: in several datasets, the k-core centrality can lead to sub-optimal performance with respect to simpler metrics such as the degree and the h-index [183, 182]. Besides, for several datasets, the eigenvector centrality [40] and LocalRank [63] significantly outperform the k-core centrality [182].

The current lack of agreement on which metric better quantifies the spreading ability of the nodes can be ascribed to two main limitations of existing studies. First, most of the proposed centrality measures do not consider the properties of the specific spreading process [40, 159, 63, 183], or they are based on analytic arguments that are valid only for specific types of networks and spreading parameters [228]. As a result, the performance of these metrics strongly depends on the network topology and on the parameters that rule the target epidemic process. Second, existing works often restrict the comparison of the metrics' performance to a limited number of parameter values [182, 228], which makes it unclear how the relative performance of the metrics depends on model parameters in the whole parameter space.

In this Chapter, we present a method to overcome both limitations. After reviewing in Section 6.1 state-of-the-art centrality measures known to be efficient in identifying influential spreaders, we define a new centrality in Section 6.2, which we call *ViralRank*, directly built on the random-walk effective distance (RWED) defined in Chapter 4. In particular, the ViralRank score of a node is defined as its average RWED to and from all the other nodes in the network. The rationale behind this definition is that an influential spreader should be able to reach and to be reached quickly from the other nodes. As the RWED quantifies with great precision the infection arrival time for any source and target node in reaction-diffusion processes, we expect the average distance to accurately quantify how well a node can reach and be reached by the other nodes. The numerical results presented in Section 6.3 show that ViralRank is the most effective metric in identifying the influential spreaders for both contact networks in the supercritical regime, as well as for reaction-diffusion spreading processes. In contact networks, if the transmission probability is sufficiently large, ViralRank is systematically the best metric to quantify the spreading ability of a node. We provide evidence that, differently from what was previously stated [159, 228], values of the transmission probability well above the critical point are relevant values to real diseases and computer viruses. For network-driven reaction-diffusion processes, ViralRank is the best-performing metric for almost all the analyzed parameter values. Besides, we show analytically that ViralRank can be

written in terms of the classical Friedkin-Johnsen social influence model, introduced in [114] and recently used to predict individuals' final opinions in controlled experiments [113, 112]. We also show that the famous PageRank centrality (2.32) can be interpreted as the average of a specific function built on the network RWED. Our findings demonstrate that the RWED can be used, in addition to estimate the infection arrival time, to quantify the nodes' spreading ability significantly better than existing metrics, bringing us closer to the optimal solution to the problem of identifying the influential spreaders for both contact-network and reaction-diffusion processes. The results presented in this Chapter are discussed in [149].

6.1. State-of-the-art centrality measures

In this Section we define existing state-of-the-art metrics known to be competitive for the identification of influential spreaders. In addition to the degree (or strength for weighted networks) and k-core centrality, that are all defined in Section 2.1.2, we select three state-of-the-art metrics: the random-walk accessibility a_i [85], LocalRank l_i [63] and the non-backtracking centrality n_i [192]. All metrics considered here are parameter-independent topological quantities build directly from the networks adjacency matrix $\mathbf{A}$. Contrary, the new metric that we introduce in Section 6.2 has the possibility for tuning of an external parameter that depends on the reaction-diffusion spreading dynamics.

Random-walk accessibility, a. The (generalized) *random-walk accessibility* is a measure that quantifies the diversity of access of individual nodes via random walks. The accessibility is defined by the exponential of the Shannon entropy [85]

$$a_i = \exp\left(-\sum_j [\exp(\mathbf{P})]_{ij} \ln[\exp(\mathbf{P})]_{ij}\right). \tag{6.1}$$

Here P_{ij} is the transition matrix, which is obtained by normalizing the unweighted adjacency matrix A_{ij} by the degree $k_i = \sum_l A_{il}$, or the weighted adjacency matrix W_{ij} by the strength $s_i = \sum_l W_{il}$ for weighted networks. The exponential of the transition matrix $\exp(\mathbf{P})$ is used so that longer walks are penalized. The accessibility has proven to identify influential spreaders particularly well in geographically embedded networks.

LocalRank, l. *LocalRank* is a centrality that generalizes the degree (2.4) by considering the nearest and the next-to-nearest neighbors to fourth order [63]. It is defined for node i as

$$l_i = \sum_k A_{ik} \sum_m A_{km} \sum_n A_{mn}(1 + \sum_r A_{nr}). \tag{6.2}$$

Thus, LocalRank gives high score to nodes that have higher degree neighbors to third and fourth order. This metric has been shown to be competitive to identify influential spreaders in [182].

Non-backtracking centrality, n. The *non-backtracking centrality* [192] can be introduced to overcome the limitation of the eigenvector centrality (2.9), by considering the Hashimoto or non-backtracking matrix [135, 169]. Given an undirected network with E edges, we construct a directed version of it with $2E$ edges, where each original edge has been replaced by two directed edges pointing in opposite directions. The non-backtracking matrix $\mathbf{B}$ is the $2E \times 2E$ non-symmetric matrix, where each element corresponds to a pair of directed edges, defined as

$$B_{i\to j,k\to l} = \delta_{jk}(1-\delta_{il}). \tag{6.3}$$

Thus the only non-zero elements of $\mathbf{B}$ are the ones defining non-backtracking paths of lengths two, from i to l, via j, with $j = k$ and $l \neq i$. Using (6.3), the non-backtracking centrality is defined as

$$n_i = \sum_j A_{ij} b_{i\to j}, \tag{6.4}$$

where $b_{i\to j}$ is the eigenvector of the non-backtracking matrix $\mathbf{B}$ corresponding to the largest eigenvalue. For the Perron-Frobenius theorem [283], the largest eigenvalue of $\mathbf{B}$ is always real and positive and the the same holds for the components of the corresponding eigenvector $b_{i\to j}$, which assures that $n_i > 0$, $\forall i$. A much faster calculation of n_i can be carried out via the Ihara-Bass determinant as the first N elements of the leading left eigenvector of the $2N \times 2N$ matrix [169]

$$\mathbf{B}' = \begin{pmatrix} \mathbf{0} & \mathbf{K}-\mathbf{I} \\ -\mathbf{I} & \mathbf{A} \end{pmatrix} \tag{6.5}$$

where $K_{ij} = \delta_{ij}k_i$ is the diagonal matrix with the degrees k_i as entries and $(\mathbf{I})_{ij} = \delta_{ij}$ is the identity matrix. Radicchi *et al.* [228] showed that the non-backtracking centrality is the most competitive metric to identify influential spreaders at criticality.

6.2. ViralRank

6.2.1. Interpretation and small λ expansion

Previous works [145, 120, 121, 46] have pointed out that in order to predict the hitting time of a spreading process in geographically-embedded systems, the network topology and the corresponding weight flows play a more fundamental role than the geographical distance. The main idea behind ViralRank is to give a score to the nodes based on

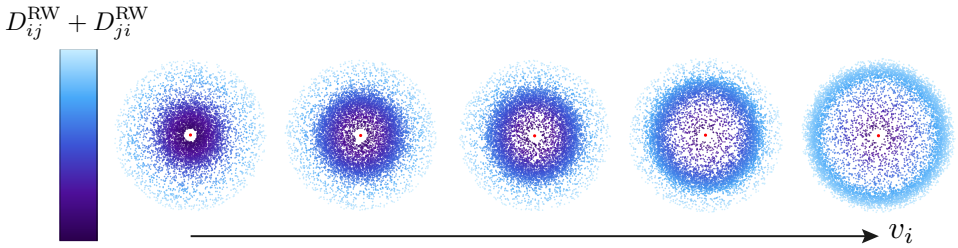


Figure 6.1: Illustration of the ViralRank centrality v_i in terms of the RWED D_{ij}^{RW} for different seed nodes i (the central red points in the figure). The clouds of nodes around each given seed node i represent the other nodes $\{j\}$ in the network. Their graphical distance from the center of the cloud is proportional to their total RWED $(D_{ij}^{\mathrm{RW}} + D_{ji}^{\mathrm{RW}})$ from the source node i; their color ranges from dark-blue (low distance) to white (high distance). The average value of all distances yields the ViralRank score v_i (horizontal axis). The cases depicted here represent examples of source nodes i with from a low ViralRank score node (left) with the majority of the other nodes grouping around the central node at low radius, to a high ViralRank score (right) defined by most nodes belonging to the peripheral sector of effective distances.

the RWED $D_{ij}^{\mathrm{RW}}(\lambda)$ defined by (4.22) which quantifies almost perfectly the hitting time of reaction-diffusion processes on networks. Importantly, the calculation of $D_{ij}^{\mathrm{RW}}(\lambda)$ is computed directly from the network adjacency matrix A_{ij}, whereas λ is the parameter defined by (4.12) that allows one to input specific epidemic and mobility rates.

We define the ViralRank score of a node i as the average RWED from and to all other nodes in the network

$$v_i(\lambda) = \frac{1}{N} \sum_j \left(D_{ij}^{\mathrm{RW}}(\lambda) + D_{ji}^{\mathrm{RW}}(\lambda) \right). \tag{6.6}$$

The nodes are therefore ranked in order of *increasing* ViralRank score: a node is central if it has, on average, small effective distance from and to the other nodes in the network[1]. As the nodes ranked high by ViralRank tend to have small effective distance from the other nodes, we expect them to generate larger epidemic outbreaks than peripheral nodes when they are chosen as the seed of a spreading process, see Figure 6.1.

Remarkably, (6.6) can be interpreted in an elegant way within a statistical physics picture. The RWED (4.22) can indeed be written as

$$D_{ij}^{\mathrm{RW}}(\lambda) = -\ln Z_{ij}(\lambda), \tag{6.7}$$

[1]To compare ViralRank's performance with that of metrics that rank the nodes in order of *decreasing* score (e.g., degree), we use $-v_i$. This way, the nodes are again ranked in order of decreasing (yet increasing in modulus) score. To keep the terminology simple, we always refer to the correlation between $-v_i$ and node's spreading ability as ViralRank's performance.

where

$$Z_{ij}(\lambda) = \sum_{n=1}^{\infty} e^{\ln H_{ij}(n)} e^{-\lambda n} = \langle e^{-\lambda n_{ij}} \rangle \,, \tag{6.8}$$

for $i \neq j$ is a function that sums all the random walks that start in i and end when arrive in j, while $Z_{ii}(\lambda) = 1$. In the last equation the quantity

$$H_{ij}(n) = \sum_{k \neq j} \left(\mathbf{P}^{(j)} \right)_{ik}^{n-1} p_k^{(j)}, \tag{6.9}$$

is the hitting time probability (2.47) of a random walk, with transition matrix P_{ij}, and $\mathbf{P}^{(j)}$ is the $(N-1)\times(N-1)$ sub-transition matrix with row and column j removed while $\mathbf{p}^{(j)}$ is the jth column of $\mathbf{P}$ with jth component removed. The average $\langle \ldots \rangle$ in (6.9) is taken over all random walks of length $\{n\}$, weighted by the probability $H_{ij}(n)$ that selects only those walks that terminate once j is reached, see also discussion in Section 4.2.3. The analogy with statistical physics emerges if we interpret λ as the inverse temperature of a thermodynamical system. Correspondingly, $Z_{ij}(\lambda)$ can be interpreted as the *partition function*, so that the RWED corresponds to the reduced free energy per temperature. In this picture, each walk length n in the summation of the partition function is in one-to-one correspondence with a single internal energy level; the quantity $e^{\ln H_{ij}(n)}$ is the relative weight of the configurations of energy n [144], i.e. the walks of length n that terminate at j. Additionally, since H_{ij} is a probability, the (micorcanonical) entropy $\mathcal{S}_{ij}^{\text{mic}}(n) = \ln H_{ij}(n)$ of the energy level n is the *self-information* [81] (or surprisal [259]) associated to the outcome of a random walker hitting node j for the first time after n steps starting from i. The total internal energy is then given by the average of the hitting time dampened by a decreasing exponential, i.e. $\mathcal{U}_{ij} = \partial_\lambda D_{ij}^{\text{RW}}(\lambda) = \langle n_{ij} e^{-\lambda n_{ij}} \rangle \,/\, \langle e^{-\lambda n_{ij}} \rangle$, with the partition function at the denominator. The canonical entropy is obtained as $\mathcal{S}_{ij} = \lambda \mathcal{U}_{ij} - \lambda \mathcal{F}_{ij}$, where $\lambda \mathcal{F}_{ij} = D_{ij}^{\text{RW}}(\lambda) = -\ln \langle e^{-\lambda n_{ij}} \rangle$ is the reduced free energy per temperature. Thus we find

$$\mathcal{S}_{ij} = \lambda \frac{\langle n_{ij} e^{-\lambda n_{ij}} \rangle}{\langle e^{-\lambda n_{ij}} \rangle} + \ln \langle e^{-\lambda n_{ij}} \rangle \,. \tag{6.10}$$

Using the expression (4.24) of the effective distance in terms of the cumulants $\langle n_{ij}^k \rangle_c$ of the hitting time, the small-λ expansion of node i's ViralRank score reads (up to a normalization constant)

$$v_i \underset{\lambda \to 0}{\approx} \lambda \sum_j (M_{ij} + M_{ji}) + \mathcal{O}(\lambda^2), \tag{6.11}$$

where $M_{ij} = \langle n_{ij} \rangle$ is the MFPT from i to j defined recursively by (2.51). The expansion for ViralRank (6.11) in this analogy corresponds to the high temperature expansion of

the free energy [212], averaged over the nodes. In this limit, the internal energy reduces to the MFPT, whereas the higher-order terms in the expansion (6.8) give a vanishing contribution and the entropy (6.10) is also vanishing. The truncated expansion to first order shows that in the limit $\lambda \to 0$, besides a uniform factor λ, node i's ViralRank score tends to the average MFPT over the rest of the network, i.e.

$$v_i \approx \sum_j \left(M_{ij} + M_{ji}\right). \tag{6.12}$$

In the last expression, the quantity summed over the network nodes $\{j\}$ is known as the *commuting time* $C_{ij} = (M_{ij} + M_{ji})$ [62], that is the expected time for a random walk to visit node j from i and then come back to node i. The latter is also related to the graph resistance distance [16] R^*_{ij} defined by (3.7) and appearing in the effective medium self-consistent equation (3.6), by $C_{ij} = 2ER^*_{ij}$ where E is the number of edges [165, 62].

While for reaction-diffusion processes, the interpretation of $D^{\mathrm{RW}}_{ij}(\lambda)$ as a proxy for the hitting time of the spreading process makes the parameter λ unambiguously determined by the parameters of the reaction dynamics of interest, the same is not true in general. Although for contact-networks, a clear-cut criterion to choose λ is lacking, expression (6.12) shows that in the limit $\lambda \to 0$, the ViralRank score of a given node i reduces to the average MFPT needed to reach the other nodes, plus the global MFPT [255], that is the time needed for a random walk starting in the nodes other than i to reach node i. In the following, for contact-network spreading, we therefore consider the quantity $v_i = v_i(\lambda \to 0)$, defined by (6.12), as node i's ViralRank score. With this choice, a node i is central if a random walk starting at node i is able to reach and to be reached quickly for the first-time with respect to the other nodes.

In the next sections, we show that (i) there is a mathematical relation between ViralRank and the Friedkin-Johnsen opinion formation model of social influence [114]; (ii) Google's PageRank [210] can also be expressed, as ViralRank, in terms of a specific partition function.

6.2.2. ViralRank and opinion formation models

Social influence network theory [115] is a mathematical formalization of the social process of attitude change as it unfolds in a social network of interpersonal influences. The influence network construct of the theory is the social structure of the endogenous interpersonal influences, in which group members are responding to the displayed positions of the members of the group. Social influence on attitudes can be conceptualized in terms of three theoretical constructs: (i) the network of interpersonal influences, (ii) persons' susceptibilities to these interpersonal influences, and (iii) persons' initial attitudes on an issue. These three construct mediate the effects of other variables on attitude.

In the Friedkin-Johnsen (FJ) linear model of opinion formation in social networks [114],

each group member i starts with an opinion $f_i = y_i(t = 1)$, normalized as $\sum_i f_i = 1$, and recursively updates it according to the linear iterative equation

$$\mathbf{y}(t+1) = \alpha \mathbf{U}\mathbf{y}(t) + (1-\alpha)\mathbf{f}. \tag{6.13}$$

Here $\mathbf{U}$ denotes the endogenous interpersonal influence matrix, such that $0 \leq U_{ij} \leq 1$ is the relative influence of j on i, and such that $\sum_j U_{ij} = 1$. The quantity $y_i(t)$ represents the position of group member (node) i on a given issue, i.e. his/her opinion. The parameter[2] $\alpha \in (0, 1)$ represents the individual susceptibility to interpersonal influence. The opinion update (6.13) is a generalization of the older and more famous DeGroot model $\mathbf{y}(t+1) = \mathbf{U}\mathbf{y}(t)$, that describes opinion formation towards consensus [88]. The final opinion $y_i = y_i(\infty)$ of i is linearly determined by the initial opinions f_j of all the other nodes $\{j\}$ through the linear relation $\mathbf{y}(\alpha|\mathbf{f}) = \mathbf{V}\mathbf{f}$, where $\mathbf{V}(\alpha) = (1-\alpha)(\mathbf{I} - \alpha\mathbf{U})^{-1}$. The matrix $\mathbf{V}$ can therefore be interpreted as the total interpersonal effects matrix [111]. In the following, we set $\mathbf{U} = \mathbf{P}$, i.e. we assume that the interpersonal influence is completely determined by the transition matrix $P_{ij} = A_{ij}/k_i$, or $P_{ij} = W_{ij}/s_i$ for weighted networks. Interestingly, families of centrality measures can be constructed from the matrix $\mathbf{V}$. An important one, referred to as *total effects centrality* by Friedkin [111], defines node i's score as $x_j = \sum_i V_{ij}/N$. As V_{ij} represents the total interpersonal influence of j on i, x_j represents the average effect of node j on the other nodes. Interestingly, as pointed out by Friedkin and Johnsen [116], in the case of interest here ($\mathbf{U} = \mathbf{P}$), this metric is exactly equivalent to Google's PageRank [44].

In Appendix B we show that, also ViraRank can be compactly written in terms of the FJ opinion formation model as

$$v_i(\lambda) = -\frac{1}{N}\sum_j \ln\left(y_i^{(j)}(e^{-\lambda}|\mathbf{f}^{(j)})\, y_j^{(i)}(e^{-\lambda}|\mathbf{f}^{(i)})\right), \tag{6.14}$$

where $\mathbf{y}^{(j)}$ is the $(N-1)$-dimensional vector of opinions obtained by removing the jth opinion from $\mathbf{y}$ and

$$\mathbf{f}^{(j)} = \frac{e^{-\lambda}}{(1-e^{-\lambda})}\mathbf{p}^{(j)}, \tag{6.15}$$

is the $(N-1)$-dimensional vector of initial opinions that is proportional to the jth column of the transition matrix $\mathbf{P}$ with element j removed, $\mathbf{p}^{(j)}$.

The FJ opinion-formation process that leads to $y_i^{(j)}$ can be interpreted as follows: each node i starts with an "opinion" proportional to $p_i^{(j)} = P_{ij}$ (with $i \neq j$) which represents the random-walk probability of jumping from i to j in one time step. Each node iteratively updates its score by summing the probabilities P_{im} of its neighbors,

[2] In the "standard model" proposed by Friedkin and Johnsen, α is a diagonal matrix of susceptibilities set to $\alpha = 1 - U_{ii}$.

j excluded, based on the FJ dynamics; the stationary state of this iterative process is $y_i^{(j)}$ which can be therefore interpreted as a (network-determined) *effective transition probability* P_{ij}. The ViralRank score v_i therefore depends on all its effective transition probabilities $y_i^{(j)}$ and $y_j^{(i)}$.

6.2.3. The relation with Google's PageRank

The PageRank score of a node is essentially a measure of how easy it is to reach a node with a random walk that is allowed to teleport. It is thus tempting to try to recover the PageRank vector of scores by modifying the RWED in order to make it a measure of the reachability of a node for a diffusion process initiated by another node. Interestingly, a PageRank vector of scores with non-uniform teleportation [171] can be obtained by averaging a modification of the partition function (6.8).

The PageRank vector is defined as the stationary density of a random walk in discrete time on a graph, whose evolution is described by the master equation (2.32). The stationary solution $\mathbf{x} = \mathbf{x}(\infty)$ reads [124]

$$\mathbf{x} = \left(\mathbf{I} - \alpha \mathbf{P}^T\right)^{-1} (1-\alpha)\mathbf{g}, \tag{6.16}$$

where $\alpha \in (0,1)$ is the damping parameter, i.e. the probability to not randomly teleport, $\mathbf{g}$ is the preference vector normalized to unity[3] ($\sum_i g_i = 1$), and P_{ij} is the row-stochastic transition matrix of the graph (2.28). In the most commonly used version of PageRank, $g_i = 1/N$, $\forall i$, is the uniform distribution vector and the damping parameter is set to $\alpha = 0.85$. Variants that consider a node dependent preference vector have been considered in [171].

In Appendix B we show that, the PageRank equation can be obtained by averaging over the source nodes $\{i\}$ a modification of partition function (6.8) that sums all random walks including those which cross multiple times the target nodes $\{j\}$. This yields the PageRank stationary density (6.16) in the form

$$\widetilde{\mathbf{x}} = \left(\mathbf{I} - e^{-\lambda} \mathbf{P}^T\right)^{-1} (1-e^{-\lambda})\widetilde{\mathbf{g}}, \tag{6.17}$$

with dumping parameter $\alpha = e^{-\lambda}$ and "smart" preference vector [171]

$$\widetilde{g}_k = \frac{e^{-\lambda}}{(1-e^{-\lambda})} \frac{1}{N} \sum_i P_{ik}. \tag{6.18}$$

By contrast, ViralRank is based on the RWED that is the logarithm of a partition function that only includes the walks that terminate once they hit the target.

[3] This assures that also $\mathbf{x}$ is normalized to unity.

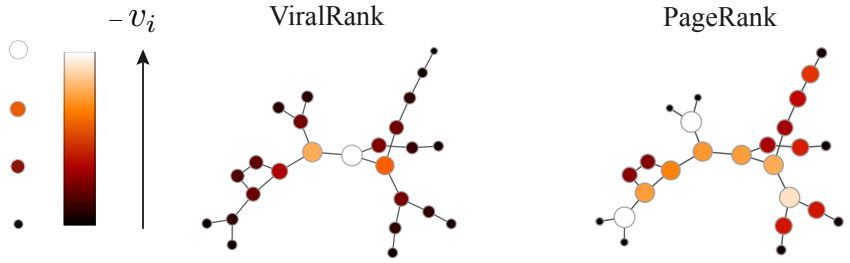


Figure 6.2: A comparison between ViralRank (6.12) and PageRank with standard dumping parameter $\alpha = 0.85$ and uniform teleportation, for a toy small-world network [274] with $N = 25$ nodes. The network is built from a ring topology where each node has $\langle k \rangle = 5$ neighbors, and by rewiring each edge with probability $p = 0.5$, as described in Section 2.1.3. The size of each node is proportional to the value of the corresponding score normalized by the maximum score in the network, and the color scale changes accordingly.

Our analytic derivations reveal the main differences between ViralRank and PageRank: (i) differently from the ViralRank score, the PageRank score does not depend logarithmically on its partition function, but linearly. This means that if a seed node i is far from a node j in the network, this will result in a small positive contribution to node i's PageRank score; by contrast, it will result in a large contribution (penalization) to its ViralRank score, proportional to D_{ij}^{RW}. (ii) The specific partition function used by PageRank also includes the walks that hit several times the arrival nodes, which results in a poor estimate of the hitting time of spreading processes. These two factors can impair PageRank's ability to identify central nodes. We qualitatively show this by analyzing a toy small-world WS network with a clear distinction between central and peripheral nodes, see Figure 6.2. The PageRank score gives a comparable score to peripheral nodes, located at the end of a branch, and central nodes, whereas ViralRank is able to clearly identify central nodes.

6.3. Identification of influential spreaders

In this Section we perform the quantitative analysis to validate the newly introduced centrality ViralRank, by considering both contact-networks and metapopulations. For contact networks we find that, among the existing metrics, there is no universally best-performing metric; however for all the analyzed networks, above a dataset-dependent threshold β_u of the transmission probability, ViralRank outperforms all the other metrics. For the metapopulation model instaead, we find that ViralRank always outperforms all the other metrics in virtually the whole parameter space.

	N	E	D	$\langle C \rangle$	$\langle k \rangle$	$\langle k^2 \rangle$	$\widetilde{\beta}_c$
ER	100	217	6	0.0509	4.34	22.44	0.2398
WS	100	300	5	0.1217	6.00	38.60	0.1840
BA	100	291	4	0.1993	5.82	62.94	0.1019

Table 6.1: Properties of the artificially constructed networks: ER with edge-creation probability $p = 0.04$, WS with $\langle k \rangle = 6$ neighbors and edge-rewiring probability $p = 0.5$ and BA with $m = 3$ new edges per time step. The different columns are: the number of nodes and edges N and E, the diameter D, the global clustering $\langle C \rangle$, the first and second moment of the degree distribution $\langle k \rangle$ and $\langle k^2 \rangle$ and the epidemic threshold $\widetilde{\beta}_c$ defined by (2.78).

6.3.1. Synthetic contact networks

We first consider artificially constructed networks with SIR dynamics. In contrast to the metapopulation approach adopted in Chapter 3 and Chapter 4, the spreading agent is directly transmitted from an infected node to its susceptible neighbors with a given probability. Each of the N nodes can be in one of three states: susceptible (S), infected (I), or recovered (R), so that the network is compartmentalized as $N = S(t)+I(t)+R(t)$. In a time step, each infected node i can infect each of its k_i neighbors with transmission probability β, and infected nodes are removed from the dynamics with probability μ. The epidemic process terminates when there are no more infected nodes in the network and the disease cannot propagate anymore. Note that since all real networks are finite, the process always terminates in a finite time even if the system is the *active phase* above the epidemic threshold, see Section 2.3.

We first study three characteristic network topologies using the generation models described in Section 2.1.3, each consisting of $N = 100$ nodes. The statistical properties of these networks are reported in Table 6.1. Within the degree-block approximation, i.e. assuming no degree correlations, the SIR epidemic threshold $\widetilde{\beta}$ is given by (2.78), such that for $\widetilde{\beta} > \widetilde{\beta}_c$ the spreading process affects a significant – i.e. non-vanishing in the thermodynamic limit– portion of the network. To assess the ability of each metric to capture the nodes' influence in the network, we compare the score they produce with the score of the nodes *spreading ability* [159, 182]. As for the rumor spreading ability (5.12), the SIR spreading ability of node i is defined as the average number of nodes in the removed state at the end of the process

$$q_i \equiv \langle \rho^R(\infty) \rangle_i \,. \tag{6.19}$$

Here $\rho^R(\infty) = \lim_{t\to\infty} \langle R(t) \rangle / N$ is the stationary state of the removed compartment, and the average $\langle \ldots \rangle_i$ is evaluated over 10^3 different realizations of the stochastic SIR dynamics, with the infection originating at node i.

The kernel density estimation of the correlation between the spreading ability (6.19)

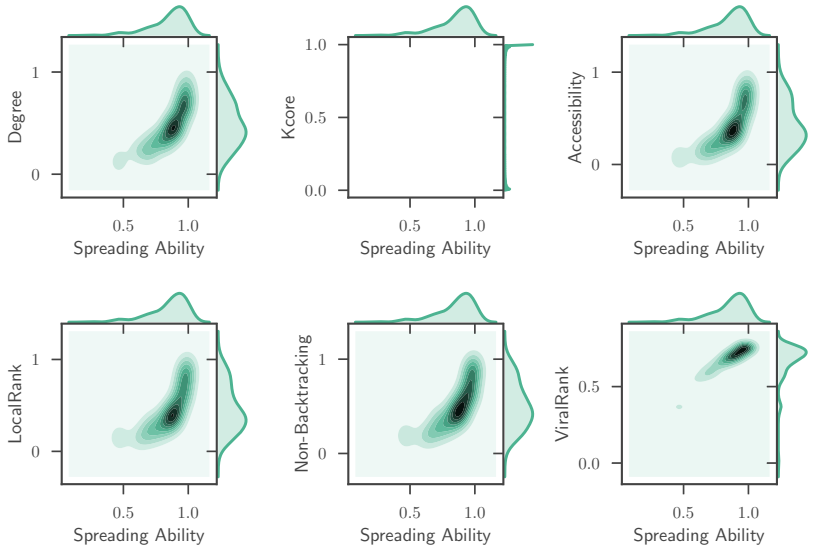


Figure 6.3: Kernel density estimation of the correlation between the max-normalized spreading ability $q/\max[q]$ and the max-normalized centralities for an ER network. The distributions are obtained at fixed ratio $\widetilde{\beta}/\widetilde{\beta}_c = 2$. Pearson correlation coefficients are respectively $r(k,q) = 0.85$, $r(k_c,q) = 0.91$, $r(a,q) = 0.83$, $r(l,q) = 0.81$, $r(n,q) = 0.83$ and $r(-v,q) = 0.96$.

and state-of-the-art centrality measures defined in Section 6.1 is shown in Figure 6.3, Figure 6.4 and Figure 6.5 for a sample point $\widetilde{\beta}/\widetilde{\beta}_c = 2$ in the parameter space (β, μ) at fixed $\mu = 1.0$. We find that for the ER and WS topologies, ViralRank gives a higher correlation with respect to all other analyzed metrics, with correlation coefficients respectively $r(-v,q) = 0.96$ and $r(-v,q) = 0.99$. For the BA network, the best performing metric at $\widetilde{\beta}/\widetilde{\beta}_c = 2$ is non-backtracking centrality with correlation coefficient $r(n,q) = 0.94$. From this preliminary result it is hard to draw a conclusion on the performance of the different metrics. We indeed expect the distance of $\widetilde{\beta}$ from $\widetilde{\beta}_c$ to significantly affect the relative metrics' performance, an aspect that is typically not extensively investigated in existing works.

To get a better insight into the relative metrics performance, we analyze a synthetic scale-free network composed of $N = 100$ nodes and $E = 189$ edges generated using the configuration model [33] with degree distribution following a power-law $\mathcal{P}(k) \sim k^{-\gamma}$, with exponent $\gamma = 2$, with an average degree $\langle k \rangle = 3.78$. To uncover how the network topology affects the metrics' performance, we replace a fraction p of its edges with edges that connects pairs of randomly selected nodes. In this way, we move continuously from a scale-free network ($p = 0$) to a random (Poissonian) topology ($p = 1$). The epidemic threshold for the scale-free network, as computed within a degree-block approximation, is $\widetilde{\beta}_c(p = 0) = 0.1412$. As the degree fluctuations decrease, the epidemic threshold

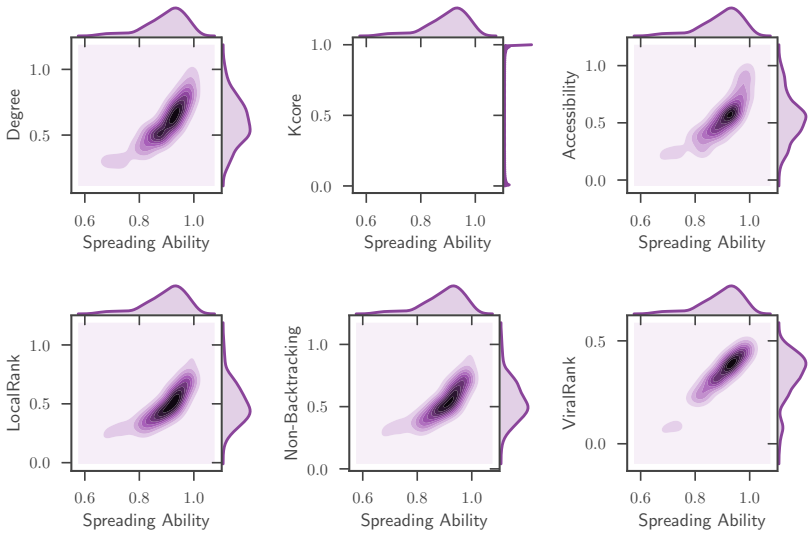


Figure 6.4: Kernel density estimation of the correlation between the max-normalized spreading ability $q/\max[q]$ and the max-normalized centralities for a WS network. The distributions are obtained at fixed ratio $\widetilde{\beta}/\widetilde{\beta}_c = 2$. Pearson correlation coefficients are respectively $r(k,q) = 0.91$, $r(k_c,q) = 0.72$, $r(a,q) = 0.88$, $r(l,q) = 0.88$, $r(n,q) = 0.88$ and $r(-v,q) = 0.99$.

increases by subsequently rewiring the edges up to $\widetilde{\beta}_c(p=1) = 0.2755 \approx 1/\langle k \rangle$ for the Poissonian random network, see Table 6.2.

In the left panel of Figure 6.6 we show the Pearson correlation $r(\cdot, q)$ between nodes' spreading ability q_i and node centralities as a function of the shuffling probability p, for a fixed value of the ratio $\widetilde{\beta}/\widetilde{\beta}_c = 4$. We find that, all metrics but the k-core and ViralRank centrality decrease their correlation with the spreading ability as the network topology becomes more homogeneous (i.e., as p increases). This reflects the fact that for a random but homogeneous topology ($p = 1$), the spreading ability spans a narrower range of values and, as a consequence, it becomes increasingly harder for the metrics to accurately estimate q_i. ViralRank is the best performing metric for all the p values; nevertheless, we shall see in the following that the metrics' relative performance critically depends on $\widetilde{\beta}$.

The right panel of Figure 6.6 shows the correlation $r(\cdot, q)$ as a function of $\widetilde{\beta}/\widetilde{\beta}_c$ for the scale-free network ($p = 0$). First, we note that around the critical point $\widetilde{\beta} \approx \widetilde{\beta}_c$, all l_i, n_i and a_i display a peak of maximum correlation with the spreading ability. This is in qualitative agreement with the fact that the non-backtracking centrality n_i is expected to match exactly for locally tree-like graphs the size of the percolating cluster at criticality with bond percolation probability $p = \widetilde{\beta}/(1+\widetilde{\beta})$ [228]; at the same time, it remains interesting that l_i and a_i display a similar behavior. This result also shows

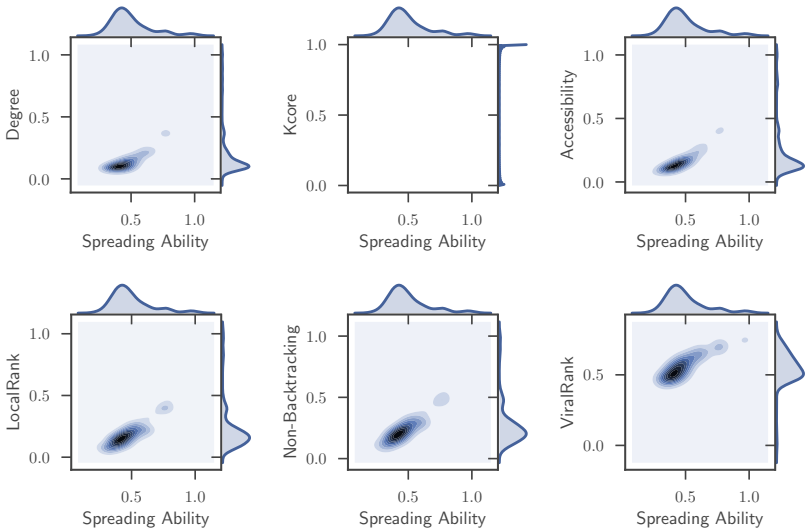


Figure 6.5: Kernel density estimation of the correlation between the max-normalized spreading ability $q/\max[q]$ and the max-normalized centralities for a BA network. The distributions are obtained at fixed ratio $\widetilde{\beta}/\widetilde{\beta}_c = 2$. Pearson correlation coefficients are respectively $r(k,q) = 0.89$, $r(k_c,q) = 0.17$, $r(a,q) = 0.92$, $r(l,q) = 0.91$, $r(n,q) = 0.94$ and $r(-v,q) = 0.85$.

that above the critical point $\widetilde{\beta}_c$, there exists an *upper-critical treshold* $\widetilde{\beta}_u > \widetilde{\beta}_c$, such that ViralRank is always the best performing metric for $\widetilde{\beta} \geq \widetilde{\beta}_u$. Real-data analysis shows that such point $\widetilde{\beta}_u$ exists for all the analyzed empirical datasets (see next Section). For the scale-free network we find $\widetilde{\beta}_u = 2.5\widetilde{\beta}_c$. We also note that there is a sensible decrease in the overall performance of all metrics as $\widetilde{\beta}$ increases. This reflects the fact that as we approach the saturation value $\beta = 1$, the distribution of nodes' spreading ability q becomes narrower, making it harder for the metrics to quantify q. Nevertheless, we emphasize that for values of $\widetilde{\beta}$ as large as $\widetilde{\beta} = 7\widetilde{\beta}_c$ of this synthetic network, we are still able to observe significant differences among the metrics' performance. This indicates that the influential spreaders identification in the supercritical regime is still a non-trivial problem, an aspect that will also emerge in real data.

To summarize, the results on synthetic networks show that in general the metrics' relative performance critically depends on the heterogeneity of the underlying network's topology and on the spreading parameters. The previous results also suggest that ViralRank significantly benefits from the spreading process being *supercritical*.

6.3.2. Empirical contact networks

We now turn our attention to twelve empirical networks (see Table 6.3 for a summary of their statistical properties and source) in which we simulate the SIR spreading pro-

	$p = 0.0$	$p = 0.2$	$p = 0.4$	$p = 0.6$	$p = 0.8$	$p = 1.0$
D	6	8	9	7	8	7
$\langle C \rangle$	0.0775	0.0822	0.0567	0.0272	0.0292	0.0293
$\langle k^2 \rangle$	30.56	27.33	24.00	19.44	18.41	17.68
$\widetilde{\beta}_c$	0.1412	0.1683	0.1938	0.2507	0.2651	0.2755

Table 6.2: Properties of a sample of the randomized networks consisting of $N = 100$ nodes and $E = 189$ edges, with average degree $\langle k \rangle = 3.78$. Each network is obtained by tuning the edge-rewiring probability p, starting from a scale-free network ($p = 0$) with degree distribution following the power-law $\mathcal{P}(k) \sim k^{-\gamma}$, with $\gamma = 2$. The different rows are the diameter D, the global clustering $\langle C \rangle$, the second moment of the degree distribution $\langle k^2 \rangle$ and the epidemic threshold $\widetilde{\beta}_c$.

cess. We provide below a brief description of each dataset and show the corresponding visualization in Figure 6.7:

Karate This is the "Zachary" karate-club social network. The dataset was collected from the members of a university karate club in 1977. Each node represents a member of the club, and each edge represents a tie between two members of the club.

Terrorists The terrorist network includes the terrorists (nodes) who belonged to the terroristic cell components centered around the 19 dead hijackers involved in the attacks at the World Trade Center of September 11th, 2001. Each edge identifies a social or economic interaction between two terrorists.

Dolphins This is the undirected social network of bottlenose dolphins. Each node is a dolphin (genus Tursiops) of a bottlenose dolphin community living off *Doubtful Sound*, a fjord in New Zealand. An edge indicates a frequent association, as measured by repeated observations between 1994 and 2001.

LesMis This network contains co-occurances of characters in Victor Hugo's novel "Les Misérables" [147]. A node represents a character and an edge shows that these two characters appeared in the same chapter of the the book. The edge weight indicating how often such a co-appearance occurred is set to unity in our simulations.

Email In the email network, the nodes represent employees of a mid-sized manufacturing company. Two employees are connected if they exchanged at least one email in the year 2010.

Jazz In the jazz network, the nodes represent jazz musicians, and the edges represent their recorded collaborations between 1912 and 1940.

Celegans This is the neural network of the roundworm *Caenorhabditis elegans*, which is the sole example of a completely mapped neural network. Nodes are neurons, and edges are interactions between them.

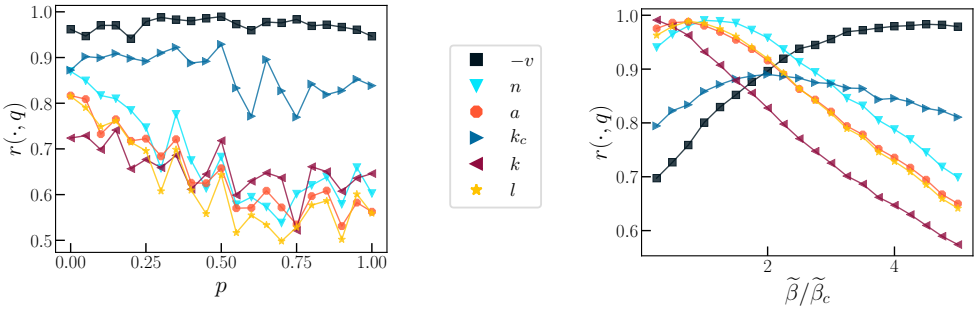


Figure 6.6: Contact-network spreading model: Correlation between nodes' centrality and nodes' spreading ability q in synthetic networks composed of 100 nodes. (left) Pearson's correlation as a function of the edges rewiring probability p, at fixed $\widetilde{\beta}/\widetilde{\beta}_c = 4$. The extreme points $p = 0$ and $p = 1$ correspond to a scale-free and to a Poissonian topology, respectively. (right) Pearson's correlation as a function of $\widetilde{\beta}/\widetilde{\beta}_c$, at fixed $p = 0$ (scale-free topology).

NetSci In the network scientists network, we select the connected giant component of the network whose nodes are the scientists working in network science. Two nodes are linked if they co-authored at least one paper including publications up until early 2006.

Flights This is the network of the 500 busiest commercial airports in the United States. An edge exists between two nodes (airports) if a flight was scheduled between them in 2002. The weights, set to unity for the simulations of contact networks, correspond to the number of seats available on each scheduled flight.

Protein The protein dataset gives the network of protein-protein interactions contained in yeast. Each node represents a protein, and an edge represents a metabolic interaction between two proteins.

Facebook In the Facebook dataset, the nodes represent Facebook users, and the edges represent their friendship relations collected from survey participants using the Facebook mobile-phone app.

PowerGrid This geographically embedded network is the high-voltage power grid located in the western United States. An edge represents a high-voltage power supply line. A node is either a generator, a transformer or a substation.

As in the previous Section, we study how the metrics' performance depends on $\widetilde{\beta}/\widetilde{\beta}_c$. In agreement with the results on synthetic networks, we find that for all the analyzed datasets, there exists a dataset-dependent value $\widetilde{\beta}_u$ such that ViralRank is the best-performing metric for $\widetilde{\beta} \geq \widetilde{\beta}_u$, see Figure 6.8. The value $\widetilde{\beta}_u$ is always larger than

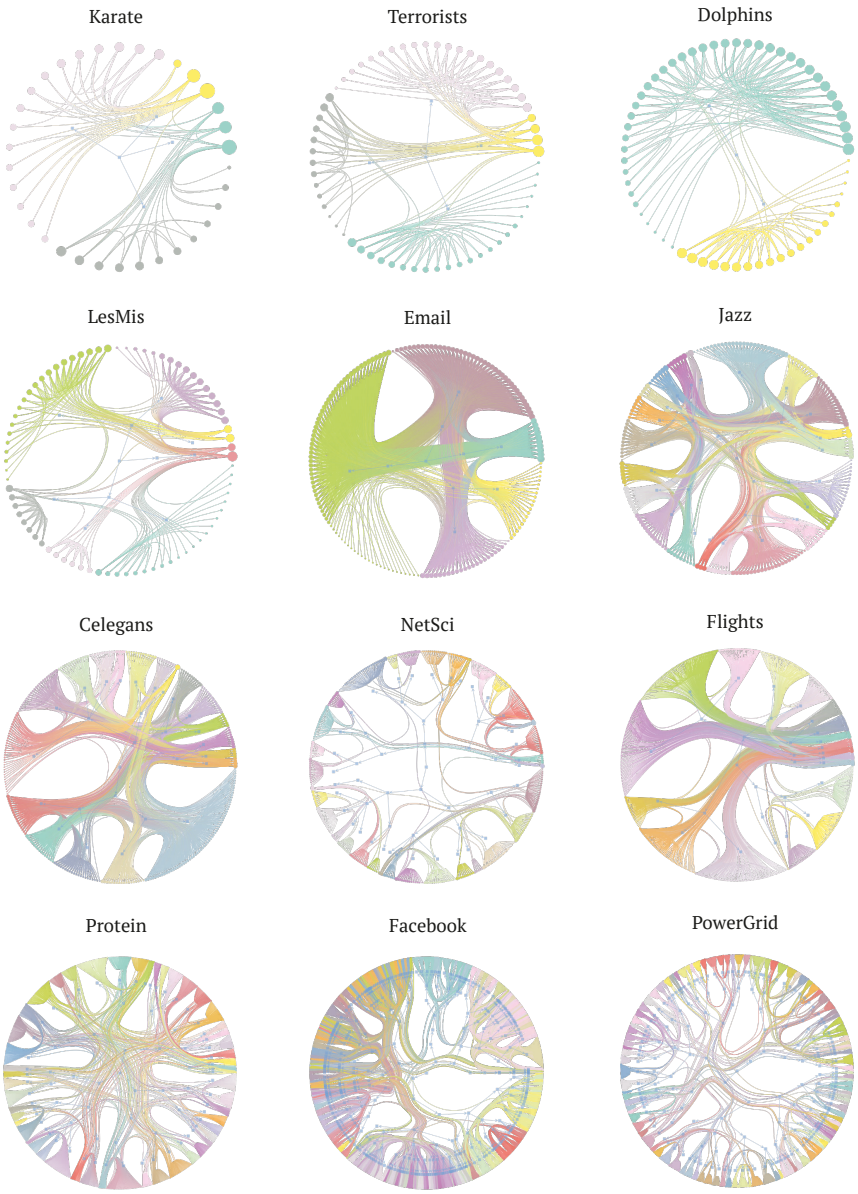


Figure 6.7: Visualization of all datasets used in the simulations (from top left): karate club friendships, 9/11 terrorists, dolphin interactions, "Les Misérables" characters co-appearances, emails, jazz collaborations, *C. elegans* neural connections, network scientists co-authorships, U.S. flights, protein interactions, Facebook friendships and U.S. power-grid supply lines. The best network partition is inferred using a multilevel Markov chain Monte Carlo algorithm [220].

	N	E	D	$\langle C\rangle$	$\langle k\rangle$	$\langle k^2\rangle$	$\widetilde{\beta}_c$	$\widetilde{\beta}_u/\widetilde{\beta}_c$	Ref.
Karate	34	78	5	0.57	4.59	35.65	0.1477	2.50	[280]
Terrorists	62	152	5	0.49	4.90	40.03	0.1396	2.50	[167]
Dolphins	62	159	8	0.26	5.13	34.90	0.1723	2.00	[184]
LesMis	77	254	5	0.57	6.60	79.53	0.0905	3.50	[166]
Email	167	3250	5	0.59	38.92	2508.78	0.0158	6.50	[195]
Jazz	198	2742	6	0.62	27.70	1070.24	0.0266	4.25	[125]
Celegans	297	2148	5	0.29	14.04	365.70	0.0399	5.75	[274]
NetSci	379	914	17	0.74	1.15	9.22	0.1424	2.00	[200]
Flights	500	2980	7	0.62	11.92	641.12	0.0189	8.00	[74]
Protein	1458	1948	19	0.07	2.08	14.85	0.1632	2.25	[79]
Facebook	4039	88234	8	0.61	43.69	4656.14	0.0095	4.75	[178]
PowerGrid	4941	6594	46	0.08	2.67	10.33	0.3483	1.50	[274]

Table 6.3: Properties of all the datasets analyzed. The different columns are the number of nodes and edges N and E, the diameter D, the global clustering $\langle C\rangle$, the first and second moment of the degree distribution $\langle k\rangle$ and $\langle k^2\rangle$ and the epidemic threshold $\widetilde{\beta}_c$; the last two columns are the upper-critical threshold $\widetilde{\beta}_u$ in units of $\widetilde{\beta}_c$ above which ViralRank outperforms all analyzed metrics and the last column the dataset source.

$\widetilde{\beta}_c$, which confirms that ViralRank is the most effective metric for the identification of influential spreaders in the *supercritical regime*, see Table 6.3. The largest ($\widetilde{\beta}_u = 8\widetilde{\beta}_c$) and smallest ($\widetilde{\beta}_u = 1.5\widetilde{\beta}_c$) values of $\widetilde{\beta}_u$ are observed for the U.S. flights and U.S. power-grid supply lines, respectively. By contrast, other metrics perform better in the vicinity of the critical point; which metric performs best in this parameter region critically depends on the considered dataset. At the critical point $\widetilde{\beta}_c$, the best performing metrics are, for almost all datasets, the non-backtracking centrality n_i and LocalRank l_i. Interestingly, for all the analyzed datasets, k-core centrality is the second-best performing metric (after ViralRank) in the supercritical regime.

These results demonstrate that among the existing metrics, there is no universally best-performing metric; the only consistent conclusion is that ViralRank outperforms all the other metrics for processes sufficiently far from criticality. Therefore, the optimal choice of a metric for ranking the nodes critically depends not only on the considered dataset, but also on the parameters of the particular spreading process in exam. Remarkably, in most of the analyzed datasets, not only ViralRank outperforms other metrics in the $\widetilde{\beta} > \widetilde{\beta}_u$ range, but it also approaches the perfect correlation with the spreading ability, $r(-v, q) = 1$, for a range of $\widetilde{\beta}$ values within the supercritical region.

While ViralRank consistently outperforms the other metrics for $\widetilde{\beta} > \widetilde{\beta}_u$, we expect its performance to dwindle as β approaches the maximum value. Indeed, for $\beta = 1$, all nodes are eventually in the recovered state for *any* initial condition and, as a result, the

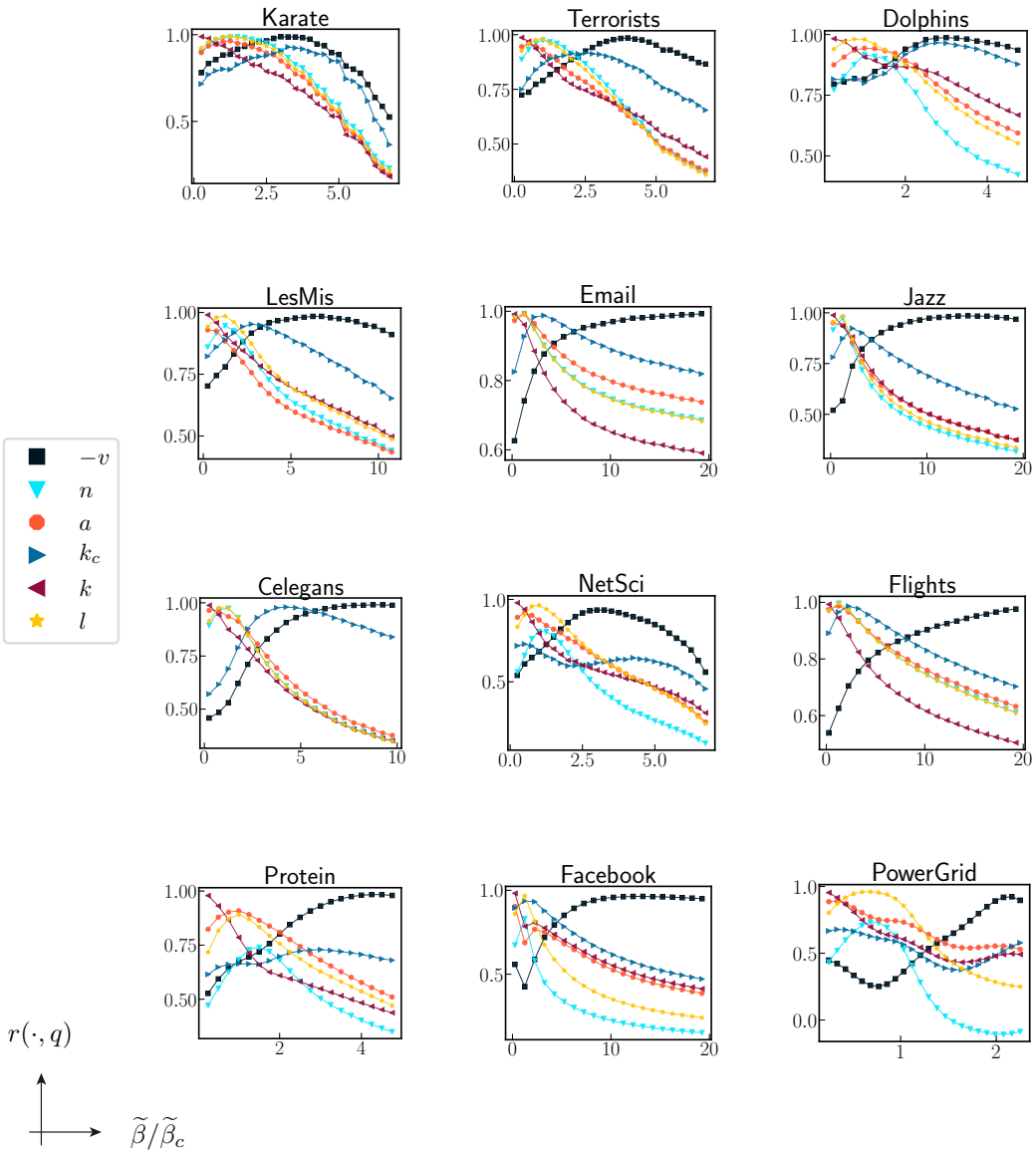


Figure 6.8: Contact-network spreading model: Comparison between nodes' centrality and nodes' spreading ability q in real networks. Pearson's correlation as a function of $\widetilde{\beta}/\widetilde{\beta}_c$ for (from top left): karate club friendships, 9/11 terrorists, dolphin interactions, "Les Misérables" characters co-appearances, emails, jazz collaborations, *C. elegans* neural connections, network scientists co-authorships, U.S. flights, protein interactions, Facebook friendships and U.S. power-grid supply lines.

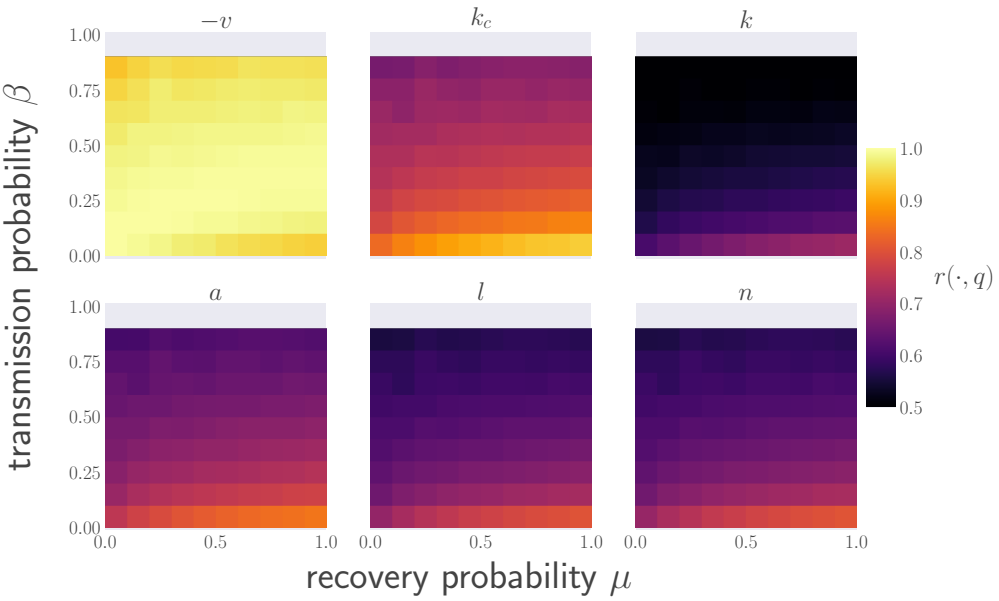


Figure 6.9: Contact-network spreading: Comparison between node centrality and node spreading ability q in the whole parameter space, for the email network ($\beta_c = 0.0158\mu$). The heat-map represents the Pearson's correlation coefficient $r(\cdot, q)$ between the nodes' centrality score and spreading ability in the (β, μ) parameter space; the colors range from black ($r = 0.5$) to yellow ($r = 1$).

nodes all have spreading ability equal to one. To quantify the extent of the parameter region over which we are able to quantify the nodes' spreading ability, we study the complete parameter space (β, μ) of transmission and recovery probabilities. We find that ViralRank is able to quantify the spreading ability for a much larger parameter region than existing metrics, see Figure 6.9. Remarkably, the correlation between ViralRank and the spreading ability is still larger than 0.95 for values of β as large as $\beta = 0.9$ and still larger than 0.90 even for $\beta = 0.99$. By contrast, for such large values of β, all the other metrics are essentially uncorrelated with q_i. Only at the saturation value $\beta = 1$ ViralRank loses its correlation with nodes' spreading ability.

The optimal performance of ViralRank for $\widetilde{\beta} > \widetilde{\beta}_u$ motivates the following question: how far are real spreading processes from criticality? To address this question, we use publicly available values of the reproductive number $\mathcal{R}_0$ of a set of real diseases. More specifically, the ranges $[\mathcal{R}_0^{min}, \mathcal{R}_0^{max}]$ of observed reproductive numbers (Table 10.2 in [17]) and publicly available values of observed transmission probabilities for computer viruses (Table 2 in [8]). We find that, by assuming the SIR dynamics on the analyzed datasets, not only real cases fall into the supercritical regime, but a number of them are in the region $\widetilde{\beta} > \widetilde{\beta}_u$ where ViralRank outperforms the other metrics in identifying influential spreaders. Below we provide the details of our analysis. For a given disease,

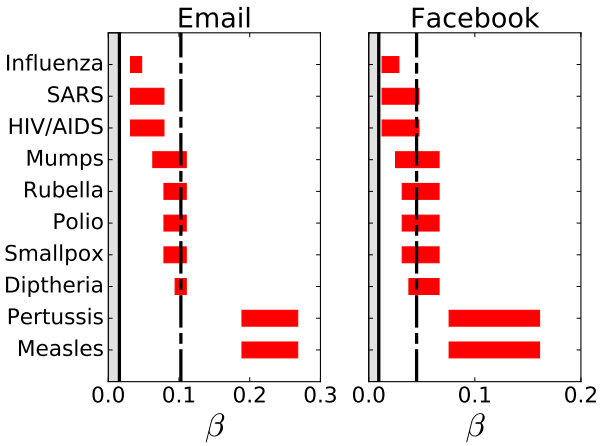


Figure 6.10: Transmission probability β corresponding to real diseases in the Email and Facebook datasets. The β ranges (red horizontal bars) match the ranges $[\mathcal{R}_0^{min}, \mathcal{R}_0^{max}]$ observed for real diseases, taken from Table 10.2 in [17]. By assuming $\mu = 1$, the $\mathcal{R}_0$ values are converted into β values according to (2.79). The continuous and dashed vertical lines represent the epidemic threshold β_c and the upper-critical point β_u such that for $\beta > \beta_u$ ViralRank is the best-performing metric.

the reproductive number $\mathcal{R}_0$ is defined as the number of secondary infections caused by an infected node in an entirely susceptible population [7]. For the SIR model the heterogeneous mean-field approximation [193] yields the *renormalized* basic reproductive number $\mathcal{R}_0 \approx (\langle k^2 \rangle / \langle k \rangle - 1)\beta/\mu$, defined by (2.79). Therefore, we can use this formula and the observed ranges $[\mathcal{R}_0^{min}, \mathcal{R}_0^{max}]$ of reproductive numbers to estimate, for each real disease and each network of interest, the expected lower and upper bounds (denoted as $\widetilde{\beta}_{min}$ and $\widetilde{\beta}_{max}$, respectively) for realistic values of $\widetilde{\beta}$. We use this procedure to estimate the interval $[\widetilde{\beta}_{min}, \widetilde{\beta}_{max}]$ for the ten diseases of Table 10.2 in [17] in two datasets, Email and Facebook. The underlying assumption is that to some extent, these two networks can be considered as proxies for the social interactions in real life through which diseases can be transmitted. We find that, for both datasets real diseases fall in the supercritical regime, and often in the region $\widetilde{\beta} > \widetilde{\beta}_u$ where ViralRank outperforms the other metrics in identifying the influential spreaders, see Figure 6.10. For example, for the Facebook dataset, the lowest $\mathcal{R}_0^{min}$ value (Influenza, SARS, HIV/AIDS, $\mathcal{R}_0^{min} = 2$) leads to $\widetilde{\beta}_{min} = 2\widetilde{\beta}_c$, which lies still below $\widetilde{\beta}_u = 4.75\widetilde{\beta}_c$. On the other hand, the upper value of $\widetilde{\beta}$ for SARS and HIV/AIDS lies above $\widetilde{\beta}_u$ ($\widetilde{\beta}_{max} = 5\widetilde{\beta}_c$). The $\widetilde{\beta}$ ranges for the diseases with the largest $\mathcal{R}_0^{min}$ (Measles, Pertussis, $\mathcal{R}_0^{min} = 12$) lie well above $\widetilde{\beta}_u$ ($\widetilde{\beta}_{min} = 12\widetilde{\beta}_c$ and $\widetilde{\beta}_{max} = 17\widetilde{\beta}_c$ for such diseases). Values of the transmission probability for some computer viruses [157, 156, 217] can be found in Table 2 from [8]. All the non-zero values reported in that table lie well above the critical point β_c (at $\mu = 1$) for the Email dataset. The *Word Macro Virus* ($\beta = 0.7$) falls in the region where ViralRank significantly outperforms the other metrics; the *Excel Macro Virus* ($\beta = 0.1$) falls below but close to the point $\beta_u = 0.103$, whereas the *Generic.exe Virus* falls in the region where the k-core centrality is the best performing metric.

These examples indicate that, by assuming a SIR dynamics, we expect the propagation of real diseases and computer viruses to be a supercritical diffusion process. We

acknowledge that our argument above is simplified, as it assumes a free propagation of the disease (i.e., no external intervention aimed at limiting the impact of the disease) on an isolated population, which is unlikely to happen in the propagation of real diseases. Nevertheless, our assumptions are the same as those of all previous studies [159, 182, 183, 228] that compared the performance of metrics based on the standard SIR model. Our argument therefore shows that, in the traditional setting for the benchmarking of metrics for the influential spreaders identification, the propagation of real diseases and computer viruses falls in the supercritical regime, and ViralRank is often the best performing metric in identifying the influential spreaders. A study of the problem in a more realistic setting goes beyond the scope of this analysis as it would require a more complex model of propagation, an accurate calibration of model parameters, and the possibility of external intervention (such as vaccination and travel restriction in the case of diseases).

6.3.3. Metapopulations

While contact networks can model epidemic spreading in a population where to each individual corresponds a node, in order to model global contagion processes in structured populations, we consider the metapopulation model, see Section 2.3.4. Now multiple individuals, of different epidemiological compartments, can only interact with individuals that are located in the same geographical location. At each time step of the dynamics, individuals can (i) interact with individuals located at the same node (*reaction*); (ii) travel across locations (*diffusion*).

In the following we assume again the SIR compartmentalization; the generalization to arbitrary complicated compartment models is obviously possible, but the SIR model often provides the sufficient level complexity necessary to describe real epidemic processes [73]. We study epidemics spreading through the U.S. air-traffic network (dataset "*Flights*" described in the previous Section): each node j is an airport to which is associated a subpopulation of size N_j; each airport is connected to the others via the weighted adjacency matrix W_{ij}. The weight W_{ij} represents the undirected flux of passengers between airport i and j per day. The probability that an individual located at node i will travel to node j at a given time is proportional to the matrix element $P_{ij} = W_{ij}/\sum_k W_{ik}$, and to the diffusion rate α, defined by (2.87). The epidemic is simulated by numerically integrating the set of equations (2.89) for $\chi = 0$.

Despite the growing interest in reaction-diffusion processes [72, 73, 75, 15, 32, 46], also spurred by their application in disease forecasting [261], the identification of influential spreaders for such dynamics has attracted less attention compared to the analogous problem for contact networks. Here, we fill this gap by comparing different centrality measures with respect to their ability to identify those subpopulations that are able to infect a large portion of the metapopulation in a short time.

As discussed in Section 2.3.4, the model defined by (2.89) is built on a Markovian

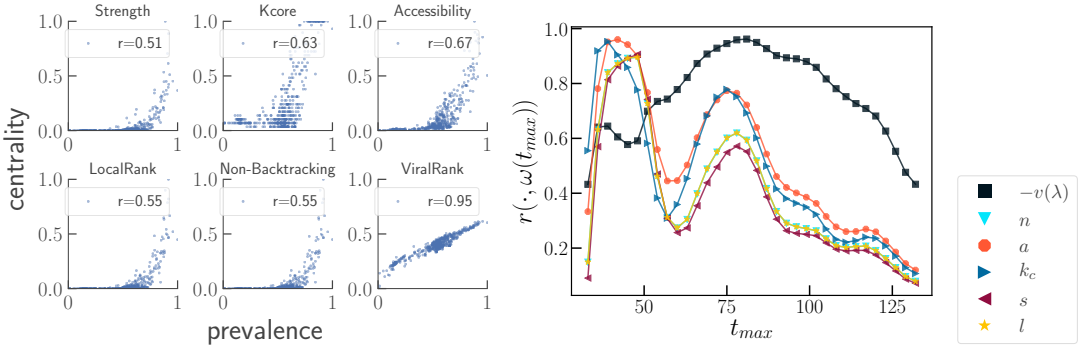
assumption and above the epidemic threshold $\widetilde{\beta}_c = 1$, all the nodes will eventually have at least one infected individual after a sufficiently long time independently on the initial condition. This however, makes it impossible to quantify the nodes' ground-truth spreading ability by measuring the asymptotic (stationary) number of subpopulations with at least one infected individual, in a similar way as we did for contact networks. To avoid this, we halt the simulations at a given threshold time t_{max}. The latter is set to half of the characteristic time for travel, given by the inverse of the diffusion rate which is set to the value $\alpha = 0.003$ d^{-1} (in unit of days), normalized by the basic reproductive number, i.e. $t_{max}(\mathcal{R}_0) = (2\mathcal{R}_0\alpha)^{-1}$. We then use the epidemic prevalence $\omega_i(t_{max})$ in the metapopulation, i.e. the number of infected subopulations at time t_{max}, for the given seed i as the initiator of the spreading process, instead of the spreading ability (6.19) as ground-truth measure of nodes relative importance.

The definition of ViralRank for contact networks (6.12) takes into account a formal limit of vanishing λ. In this limit, the ViralRank score of a node is equal to the average MFPT of a random walk. By contrast, for metapopulations, the parameter λ has a direct relation with the dynamics parameters $\mathcal{R}_0, \mu, \alpha$ given by (4.12), namely

$$\lambda(\mathcal{R}_0, \mu, \alpha) = \ln\left((\mathcal{R}_0 - 1)\frac{\mu}{\alpha}e^{-\gamma_e}\right). \tag{6.20}$$

Here, $\mathcal{R}_0 = \beta/\mu$ is the basic reproductive number in homogeneously mixed populations, μ and α the recovery and diffusion rates respectively and γ_e is the Euler-Mascheroni constant. This relation guarantees that the RWED D_{ij}^{RW} is highly correlated with the infection arrival time, see Section 4.3. As a consequence, for $\lambda = \lambda(\mathcal{R}_0, \mu, \alpha)$, the ViralRank score $-v_i(\lambda)$ is an accurate proxy for the *average arrival time* from and to a given subpopulation i. By inverting (6.20), in order to have a positive λ, which is necessary for the RWED to be well defined, we additionally require that the basic reproductive number in our simulations always satisfies the condition $\mathcal{R}_0 > 1 + \alpha/\mu e^{\gamma_e}$. However, this additional constraint only excludes a limited interval of values from our analysis; for example, when $\mu = 0.2$ d^{-1} (in unit of days), the threshold is given by $\mathcal{R}_0 \geq 1.027$, a value very close to the actual epidemic threshold $\mathcal{R}_0 = 1$.

We compare the performance of all the centrality measures introduced previously, by replacing the unweighted degree $k_i = \sum_j A_{ij}$ with the strength $s_i = \sum_j W_{ij}$, with the number of subpopulations $\omega_i(t_{max})$ that contain at least one infected individual at time t_{max}, given that the infection seed was subpopulation i. We find that, ViralRank outperforms by a great margin all the other metrics for the chosen reference point $\mathcal{R}_0 = 2$, see Figure 6.11 (a). The performance of a metric is quantified by the linear correlation between the scores it produces and $\omega(t_{max})$. All metrics besides ViralRank display a correlation smaller than $r = 0.7$, with the best performance given by the random-walk accessibility. Contrary, ViralRank displays an almost perfectly linear relation with the epidemic prevalence $\omega(t_{max})$. The comparison yields robust results for the choice of t_{max}



(a) (b)

Figure 6.11: (a) Scatter plot of the nodes' centrality scores (vertical axis) as a function of the epidemic prevalence $\omega(t_{tmax})$ (horizontal axis) at time $t_{max} = (2\alpha\mathcal{R}_0)^{-1}$ for $\mathcal{R}_0 = 2.0$ and $\alpha = 0.003$ d^{-1} (in unit of days). For each axis, the values are normalized by the maximum value. (b) Correlation coefficient between epidemic prevalence $\omega(t_{max})$ and centrality measures as a function of the observation time t_{max}. Here t_{max} is varied by keeping the value of basic reproductive number $\mathcal{R}_0 = 2.0$ fixed as well as the diffusion rate $\alpha = 0.003$ d^{-1}.

that deviates from the analytical ansatz $t_{max} = (2\alpha\mathcal{R}_0)^{-1}$, see Figure 6.11 (b).

The correlation between the scores by the centrality measures and $\omega(t_{max})$ as a function of the basic reproductive number $\mathcal{R}_0$ (with fixed $\mu = 0.2$ d^{-1}, in unit of days) is shown in Figure 6.12 (a). ViralRank is by far the best-performing metric for all the analyzed $\mathcal{R}_0$ values. The second-best performing metric is again the random-walk accessibility a_i, followed by k-core centrality. The observed performance advantage of ViralRank can be ascribed to the fact that, differently from the other metrics, it builds directly on an accurate estimate of the infection arrival time. By extending the analysis to the whole non-trivial parameter space ($\beta > \mu$), the correlation between ViralRank and the epidemic prevalence stays larger than $r = 0.8$ for a large portion of the accessible space, see Figure 6.12 (b). ViralRank is by far the best-performing metric in the whole accessible space apart from a confined region close to the epidemic threshold diagonal $\beta = \mu$, see Figure 6.12 (c). Importantly, as all real diseases reported in Table 10.2 of reference [17] have $\mathcal{R}_0 \geq 2$, they all fall into the parameter region, above the dashed line $\mathcal{R}_0 = 2$ in Figure 6.12 (b-c), where ViralRank significantly outperforms all the other metrics.

In this Chapter, we have introduced a new network centrality, called ViralRank, which quantifies the spreading ability of single nodes significantly better than existing state-of-the-art metrics for both contact-networks and reaction-diffusion supercritical spreading. To the best of our knowledge, ViralRank is the first centrality measure built on analytic estimates of random-walk hitting times. Besides, we have showed that ViralRank can be expressed in terms of the Friedkin-Johnsen opinion formation model [114]. Differently

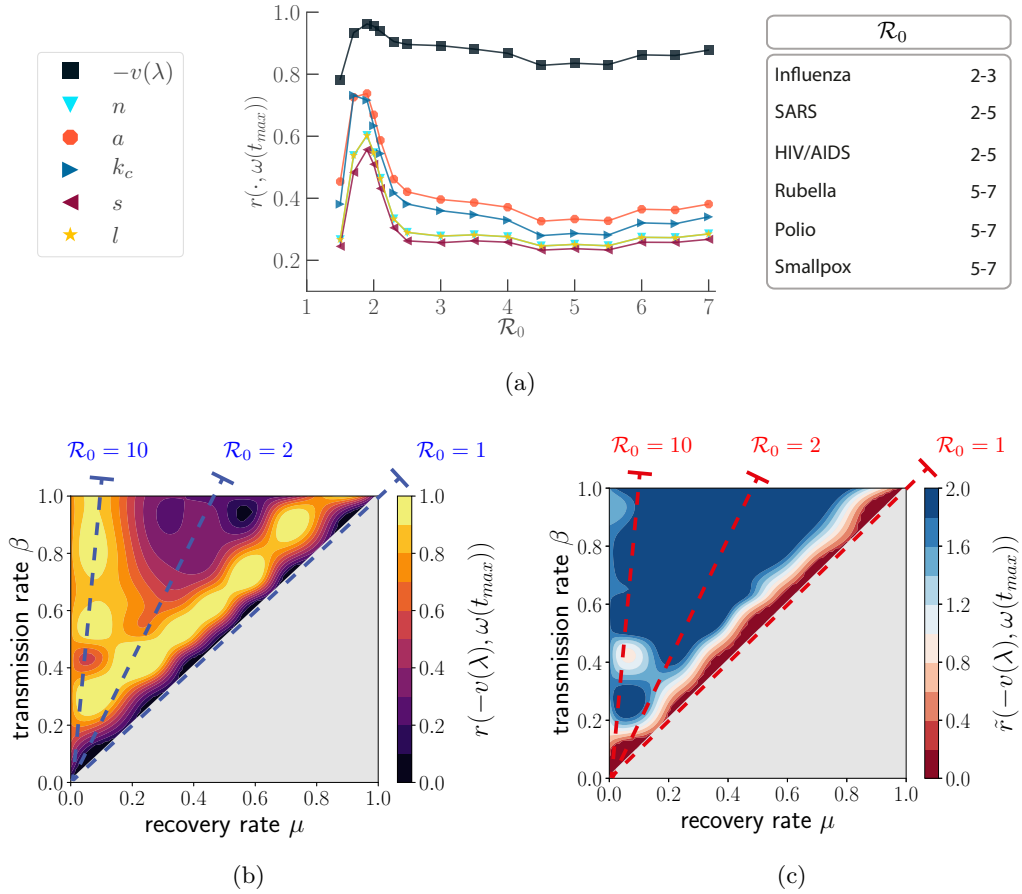


Figure 6.12: Metapopulation spreading model: A comparison between nodes' centrality and epidemic prevalence $\omega(t_{max})$ for the U.S. air-traffic network. The subpopulation strength $s_i = \sum_l W_{il}$ is used in place of the degree. (a) Pearson's correlation between nodes' centrality and $\omega(t_{max})$ as a function of the basic reproductive number $\mathcal{R}_0$, at fixed recovery rate $\mu = 0.2$ d^{-1}, in unit of days. The inset shows the known $\mathcal{R}_0$ values for some real diseases (from Table 10.2 in [17]). (b) Pearson's correlation $r(-v(\lambda), \omega(t_{max}))$ between ViralRank score and the epidemic prevalence for the non-trivial section of the accessible parameter space ($\beta > \mu$). (c) Ratio $\widetilde{r}$ between the correlations of ViralRank and the score obtained by the best performing metric (random-walk accessibility), ViralRank excluded. The dashed lines in panels (b-c) mark the lines of constant reproductive number.

from most existing studies, our analysis involved the study of the whole parameter space of the target spreading dynamics. We emphasize that, differently from the common belief, the problem of identifying the influential spreaders in the supercritical regime is important for two main reasons. First, differently from what was previously stated [159, 228], there are large differences among the metrics' performance in this regime that are revealed by our analysis. Second, and most importantly, if we assume a SIR spreading dynamics, the propagation of real diseases and computer viruses falls in the supercritical parameter region. This points out that while studying the spreading at the critical point remains an important theoretical challenge, supercritical spreading processes are in fact more likely to be of practical relevance for applications to real spreading processes.

We conclude by outlining future research directions opened by our methodology and results. It remains open to extend the RWED and ViralRank to temporal networks. This might be of extreme practical relevance inasmuch real networks exhibit strong non-Markovian effects which in turn heavily impact the properties of network diffusion processes [235, 243, 142]. Besides, the SIR model provides a realistic yet simplified model of real diseases' spreading. Extending our results to more realistic spreading models is an important direction for future research; to this end, it will be critical to calibrate the spreading simulations with the parameters observed in real epidemics. While our analysis focused on the widely-used SIR model, an extensive validation of the metrics for social contagion processes [185, 41] remains elusive, and is an important direction for future research. On this regards for rumor spreading processes, where recovery happens by direct contact and that are thus believed to be always supercritical (see Chapter 5), we expect ViralRank to significantly outperform other centralities.

Finally, ViralRank leads us closer to the optimal solution of the influential spreaders identification in the *supercritical regime*. While our results suggest that this regime is relevant for real spreading processes, it remains open to design, if at all possible, a universally best-performing metric that provides an optimal identification performance both in the supercritical and in the critical regime. For SIR type of spreading our findings confirm that the non-backtracking centrality [228] and LocalRank [182] are highly competitive around the critical point, yet their performance declines quickly in the supercritical regime, where coreness centrality gives still a good – but poorer with respect to ViralRank – estimation of influential spreading initiators. Understanding whether the effective distance can be used to build a centrality measure that is also competitive around the critical point is an intriguing challenge for future studies.

7

Conclusion

"*The problem is not to find the answer, it's to face the answer.*"

–Terence McKenna

In this thesis, we have investigated diffusion and spreading processes on networks. As a central result, we have introduced the random-walk effective distance and the centrality measure ViralRank. The aim of this work was to characterize supercritical spreading processes unfolding on networks by leveraging on random walks with absorbing states. We have shown that the resulting hitting times can provide valuable information that can be used for the characterization of spreading processes that can hardly be understood in the highly complex space of the underlying networks. In the following, we summarize the main results within each Chapter.

In Chapter 2 we presented known theoretical results with an overview on dynamical processes on complex networks, in order to have a compact reference for all subsequent Chapters. There, we laid down the theoretical framework necessary for simulating *in silico* experiments of biological and social contagions on networks. Three benchmark generation models of random networks, which extend and generalize the standard definition of discrete space defined by regular lattices, were discussed. These are the Erdös-Rényi, Watts-Strogatz and Barabási-Albert models. We used them in the subsequent Chapters as the reference models for synthetic networks that can be used as a proxy for real networks for the three key topologies: homogeneous, small-world and scale-free, respectively. Diffusion and spreading processes on networks are discussed in the broader context of non-equilibrium phase transitions. The latter are the equivalent of continuous phase transitions at equilibrium, associated with spontaneous symmetry breaking.

This is the natural setting to understand spreading processes, that are characterized by the two phases: the subcritical regime corresponding to an absorbing phase and the supercritical regime corresponding to the active phase. In this work, we focused on the characterization of supercritical spreading process, and the key results were obtained in the region of parameter space above criticality with fully deterministic dynamics. Two main classes of epidemic spreading are used: the contact-network model and the metapopulation model. The latter, directly based on reaction-diffusion equations, is used to simulate global outbreaks where each node in the network becomes a patch and is associated with an entire group of individuals in homogenous mixing between themselves.

In Chapter 3, we studied the transport properties of an ensemble of random networks with scale-free transition rates, by leveraging on effective medium theory. The latter allows for a simple and direct, yet effective approach to computing the average transition rates for the edges in random metapopulations over the disorder realizations. We use disorder to model our ignorance on the specific transportation network on which spreading might occur. This allows us to characterize the transport properties of a very general transportation network. By assuming long-range connections decaying as power laws in the lattice distance, we are able to reproduce within the disorder average, the scale-free motion empirically observed in real human mobility. Although we have used some approximations, such as a one-dimensional lattice defining the geographic space, by allowing long-range displacements we are able to map our results to realistic mobility patterns on the two-dimensional surface with Lévy flights as a proxy for the air-traffic network. Our findings demonstrate the possibility to use effective medium theory to estimate key epidemiological quantities, such as the epidemic prevalence and its exponential growth rate with high accuracy. Although only an estimate, these results open up for future research directions on reaction-diffusion spreading processes in random networks.

In Chapter 4, we have investigated the problem of first arrival for epidemic spreading in metapopulations. For this purpose we used a comprehensive dataset, not publicly available, that defines the global mobility network of air-traffic as provided by the Official Airline Guide. Each node in the corresponding network defines a subpopulation of an airport, and the edges represent the number of individuals traveling according to the corresponding weights. Then we simulated global pandemics with numerical experiments, using the metapopulation model. Each epidemic outbreak was simulated with the initial condition of a single infected individual in a given seed subpopulation, composed of otherwise susceptible individuals. We used the minimum time necessary for an infected to reach a given subpopulation as the proxy for the infection arrival time of real pandemics. Through the definition of effective distances, derived from a microscopic description of reaction-diffusion processes in a network of interconnected subpopulations, we estimated the infection arrival times. The basic definition of effective distance is derived from a dominant-path approach where a single probability-maximizing path is taken into account. By extending this definition to multiple paths and then relaxing the

assumption of direct propagation to random walks, we defined the random-walk effective distance. By comparing the numerical arrival time of the infection with the value of effective distances, the random-walk approach emerged as the most viable and accurate. Beside the theoretical challenge, our results offer a practical and near-to-optimal solution to the problem. The practical applicability of the random-walk effective distance is particularly relevant for public health issues and is a possible ideal candidate for risk assessment in real outbreak scenarios.

In Chapter 5, we modeled opinion dynamics and social contagion on the online social network Twitter. To this end, we have constructed our own dataset by downloading approximately 7 millions user posts of the political debate on Twitter regarding the Italian constitutional referendum of 2016. Two temporal networks were constructed with this procedure, one of the users mentions and one of the content retweet. We developed an analytical framework to assign dynamical opinions to users based on the content of their activity by employing machine learning techniques. The resulting time series of the global opinion averaged over all users was found in very good agreement with official pool statistics. The final result for the global opinion even outperformed official pools in estimating the real outcome of the referendum. This analysis confirms the predictive power of data analysis on Twitter in forecasting social dynamics and opinion formation in large groups of individuals. In the analysis of social contagion, we have applied a solid method based on the accessibility graph for both mentions and retweets to determine the accuracy of the respective static representations. Then we were able to use the annealed static networks as the structure on which to simulate rumor spreading between users. By using each user as the initiator of the process, we raked Twitter users in the dataset with respect to their spreading ability. Surprisingly, we found that the top spreaders are private users and not the official pro-yes and pro-no accounts, nor the official news profiles or relevant political personalities. Finally, by comparing the ranking performance of heuristic centrality measures we found that, the out-degree (number of mentions or retweets) of users reproduces the spreading ability ranking with very high accuracy outperforming all other analyzed metrics.

In Chapter 6, we have introduced a novel centrality measure called ViralRank, that predicts the stationary state of supercritical spreading processes for a given initial condition. We constructed ViralRank by averaging the random-walk effective distance defined in Chapter 4 over all source and target nodes. This corresponds to averaging over all nodes the symmetrized effective distance that defines a proper metric on a graph identifying the minimum average travel time between two nodes in both directions. By exploiting a correspondence with statistical mechanics, we found that ViralRank can be defined as the high-temperature expansion of the logarithm of a partition function that counts all walks that terminate when a target node is reached. Besides, we have found a connection between our measure and an existent opinion formation model used to predict the reach of consensus in real social experiments. Based on our definition, we also reinterpreted the famous PageRank algorithm, devised for ranking web pages

on the World Wide Web. PageRank differs from ViralRank in that, besides depending linearly, as opposed to logarithmically, on a specific partition function, it counts contributions of walks that cross multiple times a given target node. We then used ViralRank to investigate the identification of influential spreaders in real networks. We found that ViralRank systematically outperforms state-of-the-art centrality measures in correlating with the nodes' spreading ability, for both contact-network and reaction-diffusion spreading processes in the supercritical regime. By highlighting the relevance for real epidemics of supercritical spreading, our results clarify the role of random walks on networks in the identification of influential spreaders and pave the way for new research in this direction for spreading processes also far from criticality. Given the very high performance of ViralRank in the supercritical parameter region, an obvious and direct application of the measure that we have here introduced is in the context of rumor spreading processes that are believed to be always above criticality. Here, we focused on the identification of *individual* influential spreaders, in the sense that the simulated outbreaks always started from a single seed node. As recently emphasized [182], identifying a set of multiple influential spreaders might require different methods with respect to those used to identify individual influential spreaders. Extending our results to spreading processes simultaneously initiated by more than one node is a non-trivial problem for future studies, yet relevant for real-world applications (such as targeted advertising and disease immunization) where it is typically more convenient to target a large number of potential influencers (see [14] for a discussion in the context of marketing strategies).

In conclusion, this work is intended as an additional building block that adds to the theory of dynamical processes on complex networks. The intimate relation between the emergent complexity of Nature, manifested by the ubiquitous scale-free networks, and the physics of diffusion and spreading processes plays a key role in all the results obtained here. Our results showed that effective medium theory can be used efficiently to extrapolate the growth and epidemic prevalence in a very broad class of mobility networks. We have also provided evidence that random-walk hitting times can give valuable information and quantitative estimation of key epidemic quantities that can be measured in real spreading phenomena. The definition of the new centrality measure ViralRank provides an insightful and much richer interpretation of PageRank and highlights the connection between opinion formation models and the influential spreaders problem. The ubiquitous characteristics of spreading processes, especially with the recent development of online social platforms and transportation networks, makes this work relevant not only from the theoretical side but also, we believe, on the practical side.

We wish to conclude this thesis with the hope that our work, besides advancing the present state of the field of network science, will be the trigger for future studies and exciting new scientific research on diffusion and spreading processes on complex networks.

Appendix A

Lévy flights in the effective medium

Here we derive, following closely [256], the superdiffusive profile (3.24), by combining the master equation (3.11) with the power-law assumption for the transition rates (3.27). In the effective medium, the master equation (3.11) reads

$$\dot{p}_x(t) = \overline{Z} \sum_{y \in \mathbb{Z}} \frac{p_y(t) - p_x(t)}{|x-y|^{1+\alpha}}, \tag{A.1}$$

where $p_x(t)$ is the effective medium profile for free diffusion. Let us rewrite (A.1) as an infinite sum over all possible displacement lengths

$$\dot{p}_x(t) = \overline{Z} \sum_{\xi=1}^{\infty} \frac{p_{x+\xi}(t) + p_{x-\xi}(t) - 2p_x(t)}{\xi^{1+\alpha}}. \tag{A.2}$$

This equation can be solved easily in Fourier space, i.e. by multiplying e^{ikx} on both sides and sum over all x. In therms of the discrete Fourier transform $\widetilde{p}(k;t) = \sum_{x\in\mathbb{Z}} e^{ikx} p_x(t)$, we obtain

$$\begin{aligned}
\dot{\widetilde{p}}(k;t) =& \overline{Z} \sum_{x\in\mathbb{Z}} \sum_{\xi=1}^{\infty} e^{ikx} \frac{p_{x+\xi}(t) + p_{x-\xi}(t) - 2p_x(t)}{\xi^{1+\alpha}} \\
=& \overline{Z} \sum_{\xi=1}^{\infty} \frac{1}{\xi^{1+\alpha}} \left[e^{-ik\xi} + e^{ik\xi} - 2 \right] \widetilde{p}(k;t) \\
=& S(k) \widetilde{p}(k;t),
\end{aligned} \tag{A.3}$$

where

$$S(k) = \overline{Z}\left[\mathrm{Li}_{1+\alpha}\left(e^{-ik}\right) + \mathrm{Li}_{1+\alpha}\left(e^{ik}\right) - 2\mathrm{Li}_{1+\alpha}(1)\right].$$

Here $\mathrm{Li}_\nu(z) = \sum_{n=1}^{\infty} z^n/n^\nu$ is the polylogarithm function and $S(k)$ is the Fourier symbol of the transport operator defined by

$$p_x(t) = \frac{1}{2\pi}\int dk\ e^{-ikx}e^{S(k)t}. \tag{A.4}$$

Using the polylogarithm's expansion around $k = 0$ (obtained by Mathematica) one finds the following small wave-vector expression for $S(k)$

$$S(k) \sim 2\Gamma(-\alpha)\cos\left(\frac{\pi\alpha}{2}\right)\overline{Z}|k|^\alpha. \tag{A.5}$$

Note that, the sign of $S(k)$ is negative for all $\alpha \in (0, 2)$. The solution of (A.3) is given by

$$\widetilde{p}(k;t) = e^{S(k)t} \sim \exp\left(-a|k|^\alpha\right), \tag{A.6}$$

where we have used the initial condition $p_x(t = 0) = \delta_{x,0}$, which gives $\widetilde{p}(k;t) = 1$ and identified

$$a = 2\left|\Gamma(-\alpha)\cos\left(\frac{\pi\alpha}{2}\right)\right|\overline{Z}t. \tag{A.7}$$

The expansion also shows that $p_x(t)$ is asymptotically equal to a symmetric stable distribution whose Fourier transform (2.27) is exactly given by our stretched exponential (A.6) with scale parameter (A.7). Back-transforming to the position space (A.6), we obtain the asymptotic free-diffusion profile

$$p_x(t) = \mathcal{F}^{-1}\left\{e^{-a|k|^\alpha}; x\right\} \sim \frac{\Gamma(\alpha+1)}{\pi}\sin\left(\frac{\pi\alpha}{2}\right)\frac{a}{|x|^{1+\alpha}}, \tag{A.8}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform. Using this equation and the definition of a, with $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, $2\sin(y)\cos(y) = \sin(2y)$ and $\Gamma(y)\,\Gamma(-y) = \pi/\sin(\pi y)$, we recover precisely the power law (3.24) as

$$p_x(t) \sim \alpha\overline{Z}t/|x|^{1+\alpha}. \tag{A.9}$$

Appendix B

ViralRank, FJ opinion formation and PageRank

Here we derive the equations that link ViralRank with the FJ opinion formation model and with Google's PageRank. By setting $\mathbf{U} = \mathbf{P}$, the FJ model (6.13) reads

$$\begin{aligned} y_i(t+1) &= \alpha \sum_j P_{ij} y_j(t) + (1-\alpha) f_i \\ &= \frac{\alpha}{k_i} \sum_j A_{ij} y_j(t) + (1-\alpha) f_i. \end{aligned} \tag{B.1}$$

The previous equation has a simple interpretation: each node starts with an opinion f_i, and recursively updates it by averaging its neighbors' opinions. To connect the FJ model with ViralRank, it is instrumental to consider a $(N-1)\times(N-1)$ reduced matrix $\mathbf{P}^{(j)}$ obtained from $\mathbf{P}$ by removing the jth row and column. The FJ opinion formation process associated with the reduced matrix $\mathbf{P}^{(j)}$ reads

$$\begin{aligned} y_i^{(j)}(t+1) &= \alpha \sum_{m \neq j} P_{im} y_m^{(j)}(t) + (1-\alpha) f_i^{(j)} \\ &= \frac{\alpha}{k_i} \sum_{m \neq j} A_{im} y_m^{(j)}(t) + (1-\alpha) f_i^{(j)}, \end{aligned} \tag{B.2}$$

where $\mathbf{y}^{(j)}$ and $\mathbf{f}^{(j)}$ are $(N-1)$-dimensional vectors obtained by removing element j from the respective vectors. The last equation has a similar interpretation as (B.1): each node starts with an opinion $f_i^{(j)}$, and recursively updates it by considering its neighbors' opinions. Differently from (B.1), node j's opinion does not contribute to the other nodes opinions in (B.2). The stationary opinion $\mathbf{y}^{(j)}(\alpha|\mathbf{f}^{(j)})$ of the nodes is solved

by

$$\mathbf{y}^{(j)}(\alpha|\mathbf{f}^{(j)}) = (\mathbf{I}^{(j)} - \alpha\,\mathbf{P}^{(j)})^{-1}(1-\alpha)\mathbf{f}^{(j)}. \tag{B.3}$$

If $\alpha = e^{-\lambda}$, and the initial opinion vector depends on α as

$$\mathbf{f}^{(j)} = \frac{e^{-\lambda}}{(1-e^{-\lambda})}\mathbf{p}^{(j)}, \tag{B.4}$$

we finally obtain

$$y_i^{(j)}(e^{-\lambda}|\mathbf{f}^{(j)}(e^{-\lambda})) = Z_{ij}(\lambda), \tag{B.5}$$

where $Z_{ij}(\lambda)$ is the partition function (6.8) for the RWED. Thus, ViralRank centrality (6.6) can be compactly written as

$$v_i(\lambda) = -\frac{1}{N}\sum_j \ln\left(y_i^{(j)}(e^{-\lambda}|\mathbf{f}^{(j)})\, y_j^{(i)}(e^{-\lambda}|\mathbf{f}^{(i)})\right). \tag{B.6}$$

To illustrate the connection between ViralRank and PageRank, let us substitute (6.9) into the partition function (6.8), which explicitly reads

$$Z_{ij}(\mathbf{P},\lambda) = \sum_{k\neq j}\left(\mathbf{I}^{(j)} - e^{-\lambda}\mathbf{P}^{(j)}\right)^{-1}_{ik} e^{-\lambda} p_k^{(j)}. \tag{B.7}$$

Then, let us define a modified partition function

$$\widetilde{Z}_{ij}(\mathbf{P},\lambda) = \sum_{k\neq j}\left(\mathbf{I} - e^{-\lambda}\mathbf{P}\right)^{-1}_{ik} e^{-\lambda} P_{kj}. \tag{B.8}$$

Contrary to the partition function of the RWED (B.7), where only walks that terminate in j are considered, in (B.8) all walks that can also cross multiple times the target j are summed over. First, note that by rearranging the sum we obtain

$$\begin{aligned}
\widetilde{Z}_{ij}(\mathbf{P}^T,\lambda) &= \sum_{k\neq j}\sum_{n=0}^{\infty}\left(e^{-\lambda}\mathbf{P}^T\right)^n_{ik} e^{-\lambda} P^T_{kj} \\
&= \sum_{m\neq j}\sum_{n=0}^{\infty}\sum_{k\neq j} e^{-\lambda} P_{jk}\left(e^{-\lambda}\mathbf{P}\right)^{n-1}_{km} e^{-\lambda} P_{mi} \\
&= \sum_{m\neq j}\sum_{n=0}^{\infty}\left(e^{-\lambda}\mathbf{P}\right)^n_{jm} e^{-\lambda} P_{mi} \\
&= \widetilde{Z}_{ji}(\mathbf{P},\lambda).
\end{aligned} \tag{B.9}$$

By averaging the modified partition function (B.8) over the source nodes $\{i\}$ we obtain the vector $\widetilde{x}_j = N^{-1}\sum_i \widetilde{Z}_{ij}(\mathbf{P},\lambda)$ and using (B.9)

$$\widetilde{x}_j = \frac{1}{N}\sum_i \sum_{k\neq i} \left(\mathbf{I} - e^{-\lambda}\mathbf{P}^T\right)^{-1}_{jk} e^{-\lambda} P^T_{ki}. \tag{B.10}$$

Finally, if no self-loops are present $P_{ii} = 0$ and we can consider the full sum including also node i, to get precisely the PageRank score (6.16)

$$\widetilde{\mathbf{x}} = \left(\mathbf{I} - e^{-\lambda}\mathbf{P}^T\right)^{-1}(1 - e^{-\lambda})\widetilde{\mathbf{g}}, \tag{B.11}$$

where the dumping parameter is $\alpha = e^{-\lambda}$ and the "smart" preference vector [171] is defined as

$$\widetilde{g}_k = \frac{e^{-\lambda}}{(1-e^{-\lambda})}\frac{1}{N}\sum_i P_{ik}. \tag{B.12}$$

Additionally, we can impose the standard normalization by requiring that $\sum_k g_k = 1$ and rescale the whole PageRank score (B.11) by the normalization factor $(1-e^{-\lambda})/e^{-\lambda}$, so that the new quantity is correctly normalized to unity, as in (6.16).

Equation (B.11) therefore shows that PageRank with dumping parameter $\alpha = e^{-\lambda}$ and and non-uniform teleportation vector (B.12), builds on a partition function $\widetilde{Z}_{ij}$ that also includes walks that hit the target nodes $\{j\}$ multiple times. By contrast, ViralRank is built on the partition function of the RWED that only selects the walks that terminate once they hit the target.

Bibliography

[1] Official Airline Guide (OAG Ltd.): http://www.oag.com.

[2] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning From Data*, AMLBook, 2012.

[3] E. Agliari, R. Burioni, D. Cassi, and A. Vezzani, *Random walks interacting with evolving energy landscapes*, The European Physical Journal B-Condensed Matter and Complex Systems, 48 (2005), pp. 529–536.

[4] R. Albert and A.-L. Barabási, *Statistical mechanics of complex networks*, Reviews of Modern Physics, 74 (2002), p. 47.

[5] R. Albert, H. Jeong, and A.-L. Barabási, *Internet: Diameter of the world-wide web*, Nature, 401 (1999), p. 130.

[6] P. W. Anderson, *More is different*, Science, 177 (1972), pp. 393–396.

[7] R. M. Anderson, R. M. May, and B. Anderson, *Infectious Diseases of Humans: Dynamics and Control*, vol. 28, Wiley Online Library, 1992.

[8] J. L. Aron, M. O'Leary, R. A. Gove, S. Azadegan, and M. C. Schneider, *The benefits of a notification process in addressing the worsening computer virus problem: results of a survey and a simulation model*, Computers & Security, 21 (2002), pp. 142–163.

[9] N. Ashcroft and N. Mermin, *Solid State Physics*, Cengage Learning, 2011.

[10] R. Atkinson, C. Rhodes, D. Macdonald, and R. Anderson, *Scale-free dynamics in the movement patterns of jackals*, Oikos, 98 (2002), pp. 134–140.

[11] P. Bajardi, C. Poletto, J. J. Ramasco, M. Tizzoni, V. Colizza, and A. Vespignani, *Human mobility networks, travel restrictions, and the global spread of 2009 h1n1 pandemic*, PLOS One, 6 (2011), p. e16591.

[12] P. Bak, K. Chen, and C. Tang, *A forest-fire model and some thoughts on turbulence*, Physics Letters A, 147 (1990), pp. 297–300.

[13] P. Bak, C. Tang, and K. Wiesenfeld, *Self-organized criticality: An explanation of the 1/f noise*, Physical Review Letters, 59 (1987), p. 381.

[14] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, *Everyone's an influencer: quantifying influence on twitter*, in Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, ACM, 2011, pp. 65–74.

[15] D. Balcan, B. Gonçalves, H. Hu, J. J. Ramasco, V. Colizza, and A. Vespignani, *Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model*, Journal of Computational Science, 1 (2010), pp. 132–145.

[16] R. B. Bapat, *Graphs and Matrices*, vol. 27, Springer, 2010.

[17] A.-L. Barabási, *Network Science*, Cambridge University Press, 2016.

[18] A.-L. Barabási and R. Albert, *Emergence of scaling in random networks*, Science, 286 (1999), pp. 509–512.

[19] A.-L. Barabási and H. E. Stanley, *Fractal Concepts In Surface Growth*, Cambridge University Press, 1995.

[20] A. Baronchelli, A. Barrat, and R. Pastor-Satorras, *Glass transition and random walks on complex energy landscapes*, Physical Review E, 80 (2009), p. 020102.

[21] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, *The architecture of complex weighted networks*, Proceedings of the National Academy of Sciences, 101 (2004), pp. 3747–3752.

[22] A. Barrat, M. Barthelemy, and A. Vespignani, *The architecture of complex weighted networks: Measurements and models*, in Large Scale Structure And Dynamics Of Complex Networks: From Information Technology to Finance and Natural Science, World Scientific, 2007, pp. 67–92.

[23] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks*, Cambridge University Press, 2008.

[24] A. Barrat and M. Weigt, *On the properties of small-world network models*, The European Physical Journal B-Condensed Matter and Complex Systems, 13 (2000), pp. 547–560.

[25] M. Barthélemy, *Spatial networks*, Physics Reports, 499 (2011), pp. 1–101.

[26] F. Bartumeus, F. Peters, S. Pueyo, C. Marrasé, and J. Catalan, *Helical lévy walks: Adjusting searching statistics to resource availability in microzooplankton*, Proceedings of the National Academy of Sciences, 100 (2003), pp. 12771–12775.

[27] S. Battiston, J. B. Glattfelder, D. Garlaschelli, F. Lillo, and G. Caldarelli, *The Structure of Financial Networks*, Springer London, London, 2010, pp. 131–163.

[28] S. Battiston, M. Puliga, R. Kaushik, P. Tasca, and G. Caldarelli, *Debtrank: Too central to fail? financial networks, the fed and systemic risk*, Scientific Reports, 2 (2012), p. 541.

[29] F. Bauer and J. T. Lizier, *Identifying influential spreaders and efficiently estimating infection numbers in epidemic models: A walk counting approach*, EPL (Europhysics Letters), 99 (2012), p. 68007.

[30] F. Bavaud and G. Guex, *Interpolating between Random Walks and Shortest Paths: A Path Functional Approach*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 68–81.

[31] A. Békéssy, *Asymptotic enumeration of regular matrices*, Stud. Sci. Math. Hungar., 7 (1972), pp. 343–353.

[32] V. Belik, T. Geisel, and D. Brockmann, *Natural human mobility patterns and spatial spread of infectious diseases*, Physical Review X, 1 (2011), p. 011001.

[33] E. A. Bender and E. R. Canfield, *The asymptotic number of labeled graphs with given degree sequences*, Journal of Combinatorial Theory, Series A, 24 (1978), pp. 296–307.

[34] G. Bianconi and A.-L. Barabási, *Competition and multiscaling in evolving networks*, EPL (Europhysics Letters), 54 (2001), p. 436.

[35] J. Bindi, D. Colombi, F. Iannelli, N. Politi, M. Sugarelli, R. Tavarone, and E. Ubaldi, *Political discussion and leanings on twitter: the 2016 italian constitutional referendum*, arXiv preprint arXiv:1805.07388, (2018).

[36] J. J. Binney, N. J. Dowrick, A. J. Fisher, and M. E. J. Newman, *The Theory of Critical Phenomena: An Introduction to the Renormalization Group*, Oxford University Press, Inc., 1992.

[37] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, O'Reilly Media, 2009.

[38] F. Blöchl, F. J. Theis, F. Vega-Redondo, and E. O. Fisher, *Vertex centralities in input-output networks reveal the structure of modern economies*, Physical Review E, 83 (2011), p. 046127.

[39] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, *Fast unfolding of communities in large networks*, Journal of Statistical Mechanics: Theory and Experiment, 2008 (2008), p. P10008.

[40] P. Bonacich, *Factoring and weighting approaches to status scores and clique identification*, Journal of Mathematical Sociology, 2 (1972), pp. 113–120.

[41] J. Borge-Holthoefer and Y. Moreno, *Absence of influential spreaders in rumor dynamics*, Physical Review E, 85 (2012), p. 026116.

[42] A. Bovet, F. Morone, and H. A. Makse, *Predicting election trends with twitter: Hillary clinton versus donald trump*, arXiv preprint arXiv:1610.01587, (2016).

[43] L. A. Braunstein, S. V. Buldyrev, R. Cohen, S. Havlin, and H. E. Stanley, *Optimal paths in disordered complex networks*, Physical Review Letters, 91 (2003), p. 168701.

[44] S. Brin and L. Page, *The anatomy of a large-scale hypertextual web search engine*, Computer networks and ISDN systems, 30 (1998), pp. 107–117.

[45] S. R. Broadbent and J. M. Hammersley, *Percolation processes: I. crystals and mazes*, in Mathematical Proceedings of the Cambridge Philosophical Society, vol. 53, Cambridge University Press, 1957, pp. 629–641.

[46] D. Brockmann and D. Helbing, *The hidden geometry of complex, network-driven contagion phenomena*, Science, 342 (2013), pp. 1337–1342.

[47] D. Brockmann and L. Hufnagel, *Front propagation in reaction-superdiffusion dynamics: taming lévy flights with fluctuations*, Physical Review Letters, 98 (2007), p. 178301.

[48] D. Brockmann, L. Hufnagel, and T. Geisel, *The scaling laws of human travel*, Nature, 439 (2006), pp. 462–465.

[49] W. V. d. Broeck, C. Gioannini, B. Gonçalves, M. Quaggiotto, V. Colizza, and A. Vespignani, *The gleamviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale*, BMC Infectious Diseases, 11 (2011), pp. 1–14.

[50] A. D. Broido and A. Clauset, *Scale-free networks are rare*, arXiv preprint arXiv:1801.03400, (2018).

[51] D. A. G. Bruggeman, *Berechnung verschiedener physikalischer konstanten von heterogenen substanzen*, Annalen der Physik, 24 (1935), pp. 636–664.

[52] A. Bunde and S. Havlin, *Fractals and Disordered Systems*, Springer Science & Business Media, 2012.

[53] J. Byers, *The physics of data*, Nature Physics, 13 (2017), pp. 718–719.

[54] G. Caldarelli, *Scale-Free Neworks*, Oxford University Press, 2007.

[55] G. Caldarelli, A. Chessa, F. Pammolli, G. Pompa, M. Puliga, M. Riccaboni, and G. Riotta, *A multi-level geographical study of italian political elections from twitter data*, PLOS One, 9 (2014), p. e95809.

[56] C. G. Callan Jr, *Broken scale invariance in scalar field theory*, Physical Review D, 2 (1970), p. 1541.

[57] J. L. Cardy, *Field theoretic formulation of an epidemic process with immunisation*, Journal of Physics A: Mathematical and General, 16 (1983), p. L709.

[58] J. L. Cardy and P. Grassberger, *Epidemic models and percolation*, Journal of Physics A: Mathematical and General, 18 (1985), p. L267.

[59] C. Castellano, S. Fortunato, and V. Loreto, *Statistical physics of social dynamics*, Reviews of Modern Physics, 81 (2009), pp. 591–646.

[60] A. Cavagna, A. Cimarelli, I. Giardina, G. Parisi, R. Santagati, F. Stefanini, and M. Viale, *Scale-free correlations in starling flocks*, Proceedings of the National Academy of Sciences, 107 (2010), pp. 11865–11870.

[61] D. Centola and M. Macy, *Complex contagions and the weakness of long ties*, American journal of Sociology, 113 (2007), pp. 702–734.

[62] A. K. Chandra, P. Raghavan, W. L. Ruzzo, R. Smolensky, and P. Tiwari, *The electrical resistance of a graph captures its commute and cover times*, Computational Complexity, 6 (1996), pp. 312–340.

[63] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, *Identifying influential nodes in complex networks*, Physica A: Statistical Mechanics and its Applications, 391 (2012), pp. 1777–1787.

[64] K. Chen, P. Bak, and M. H. Jensen, *A deterministic critical forest fire model*, Physics Letters A, 149 (1990), pp. 207–210.

[65] X. Chen, T. Kumagai, and J. Wang, *Random conductance models with stable-like jumps i: Quenched invariance principle*, arXiv preprint arXiv:1805.04344, (2018).

[66] T. C. Choy, *Effective Medium Theory – Principles and Applications*, International Series of Monographs on Physics, Oxford University Press, New York, 1999.

[67] F. Ciulla, D. Mocanu, A. Baronchelli, B. Gonçalves, N. Perra, and A. Vespignani, *Beating the news using social media: the case study of american idol*, EPJ Data Science, 1 (2012), p. 8.

[68] A. Clauset, M. Kogan, and S. Redner, *Safe leads and lead changes in competitive team sports*, Physical Review E, 91 (2015), p. 062815.

[69] A. Clauset, C. R. Shalizi, and M. E. J. Newman, *Power-law distributions in empirical data*, SIAM Review, 51 (2009), pp. 661–703.

[70] R. Cohen and S. Havlin, *Scale-free networks are ultrasmall*, Physical Review Letters, 90 (2003), p. 058701.

[71] R. Cohen, S. Havlin, and D. Ben-Avraham, *Efficient immunization strategies for computer networks and populations*, Physical Review Letters, 91 (2003), p. 247901.

[72] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani, *The role of the airline transportation network in the prediction and predictability of global epidemics*, Proceedings of the National Academy of Sciences, 103 (2006), pp. 2015–2020.

[73] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani, *Predictability and epidemic pathways in global outbreaks of infectious diseases: the SARS case study*, BMC Medicine, 5 (2007).

[74] V. Colizza, R. Pastor-Satorras, and A. Vespignani, *Reaction-diffusion processes and metapopulation models in heterogeneous networks*, Nature Physics, 3 (2007), p. 276.

[75] V. Colizza and A. Vespignani, *Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations*, Journal of Theoretical Biology, 251 (2008), pp. 450–467.

[76] M. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, *Political polarization on twitter.*, ICWSM, 133 (2011), pp. 89–96.

[77] M. D. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini, and F. Menczer, *Predicting the political alignment of twitter users*, in 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Oct 2011, pp. 192–199.

[78] D. Coppersmith and S. Winograd, *Matrix multiplication via arithmetic progressions*, in Proceedings of the nineteenth annual ACM symposium on Theory of computing, ACM, 1987, pp. 1–6.

[79] S. Coulomb, M. Bauer, D. Bernard, and M.-C. Marsolier-Kergoat, *Gene essentiality and the topology of protein interaction networks*, Proceedings of the Royal Society B: Biological Sciences, 272 (2005), pp. 1721–1725.

[80] B. Coutinho, S. Hong, K. Albrecht, A. Dey, A.-L. Barabási, P. Torrey, M. Vogelsberger, and L. Hernquist, *The network behind the cosmic web*, arXiv preprint arXiv:1604.03236, (2016).

[81] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 2012.

[82] P. Crepey, F. P. Alvarez, and M. Barthélemy, *Epidemic variability in complex networks*, Physical Review E, 73 (2006), p. 046131.

[83] M. Cristelli, A. Tacchella, and L. Pietronero, *The heterogeneous dynamics of economic complexity*, PLOS One, 10 (2015), p. e0117174.

[84] D. J. Daley and D. G. Kendall, *Epidemics and rumours*, Nature, 204 (1964), p. 1118.

[85] G. F. De Arruda, A. L. Barbieri, P. M. Rodríguez, F. A. Rodrigues, Y. Moreno, and L. da Fontoura Costa, *Role of centrality for the identification of influential spreaders in complex networks*, Physical Review E, 90 (2014), p. 032812.

[86] C. De Dominicis and I. Giardina, *Random Fields and Spin Glasses: A Field Theory Approach*, Cambridge University Press, 2006.

[87] P.-G. De Gennes, *Scaling Concepts in Polymer Physics*, Cornell University Press, 1979.

[88] M. H. DeGroot, *Reaching a consensus*, Journal of the American Statistical Association, 69 (1974), pp. 118–121.

[89] D. del Castillo-Negrete, B. Carreras, and V. Lynch, *Front dynamics in reaction-diffusion systems with levy flights: a fractional diffusion approach*, Physical Review Letters, 91 (2003), p. 018302.

[90] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, *The spreading of misinformation online*, Proceedings of the National Academy of Sciences, 113 (2016), pp. 554–559.

[91] E. W. Dijkstra, *A note on two problems in connexion with graphs*, Numerische Mathematik, 1 (1959), pp. 269–271.

[92] P. Domingos and M. Richardson, *Mining the network value of customers*, in Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2001, pp. 57–66.

[93] S. N. Dorogovtsev, A. V. Goltsev, and J. F. Mendes, *Critical phenomena in complex networks*, Reviews of Modern Physics, 80 (2008), p. 1275.

[94] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, *K-core organization of complex networks*, Physical Review Letters, 96 (2006), p. 040601.

[95] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, *Structure of growing networks with preferential linking*, Physical Review Letters, 85 (2000), p. 4633.

[96] P. G. Doyle and J. L. Snell, *Random walks and electric networks*, vol. 22, Mathematical Association of America, 1984.

[97] K. T. Eames and M. J. Keeling, *Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases*, Proceedings of the National Academy of Sciences, 99 (2002), pp. 13330–13335.

[98] Y.-H. Eom, M. Puliga, J. Smailović, I. Mozetič, and G. Caldarelli, *Twitter-based analysis of the dynamics of collective attention to political parties*, PLOS One, 10 (2015), p. e0131184.

[99] P. Erdős and A. Hajnal, *On chromatic number of graphs and set-systems*, Acta Mathematica Academiae Scientiarum Hungarica, 17 (1966), pp. 61–99.

[100] P. Erdös and A. Rényi, *On random graphs, i*, Publicationes Mathematicae (Debrecen), 6 (1959), pp. 290–297.

[101] S. Erlander and N. F. Stewart, *The gravity model in transportation analysis: theory and extensions*, vol. 3, Vsp, 1990.

[102] O. Feinerman, I. Pinkoviezky, A. Gelblum, E. Fonio, and N. S. Gov, *The physics of cooperative transport in groups of ants*, Nature Physics, (2018), p. 1.

[103] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1, John Wiley & Sons, 1968.

[104] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 2, John Wiley & Sons, 1971.

[105] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, *The rise of social bots*, Communications of the ACM, 59 (2016), pp. 96–104.

[106] A. Fick, *On liquid diffusion*, Journal of Membrane Science, 100 (1995), pp. 33–38.

[107] R. A. Fisher, *The wave of advance of advantageous genes*, Annals of Human Genetics, 7 (1937), pp. 355–369.

[108] S. Fortunato, A. Flammini, and F. Menczer, *Scale-free network growth by ranking*, Physical Review Letters, 96 (2006), p. 218701.

[109] L. C. Freeman, *A set of measures of centrality based on betweenness*, Sociometry, (1977), pp. 35–41.

[110] M. I. Freidlin, *Functional Integration and Partial Differential Equations*, vol. 109, Princeton University Press, 2016.

[111] N. E. Friedkin, *Theoretical foundations for centrality measures*, American Journal of Sociology, 96 (1991), pp. 1478–1504.

[112] N. E. Friedkin and F. Bullo, *How truth wins in opinion dynamics along issue sequences*, Proceedings of the National Academy of Sciences, 114 (2017), pp. 11380–11385.

[113] N. E. Friedkin, P. Jia, and F. Bullo, *A theory of the evolution of social power: Natural trajectories of interpersonal influence systems along issue sequences*, Sociological Science, 3 (2016), pp. 444–472.

[114] N. E. Friedkin and E. C. Johnsen, *Social influence and opinions*, The Journal of Mathematical Sociology, 15 (1990), pp. 193–206.

[115] N. E. Friedkin and E. C. Johnsen, *Social Influence Network Theory: A Sociological Examination of Small Group Dynamics*, vol. 33, Cambridge University Press, 2011.

[116] N. E. Friedkin and E. C. Johnsen, *Two steps to obfuscation*, Social Networks, 39 (2014), pp. 12–13.

[117] H. Frisch and J. Hammersley, *Percolation processes and related topics*, Journal of the Society for Industrial and Applied Mathematics, 11 (1963), pp. 894–918.

[118] A. Gabel and S. Redner, *Random walk picture of basketball scoring*, Journal of Quantitative Analysis in Sports, 8 (2012).

[119] G. Gallavotti, *Statistical Mechanics: A Short Treatise*, Springer Science & Business Media, 2013.

[120] A. Gautreau, A. Barrat, and M. Barthélemy, *Arrival time statistics in global disease spread*, Journal of Statistical Mechanics: Theory and Experiment, 2007 (2007), p. L09001.

[121] A. Gautreau, A. Barrat, and M. Barthelemy, *Global disease spread: statistics and estimation of arrival times*, Journal of Theoretical Biology, 251 (2008), pp. 509–522.

[122] T. Geisel, J. Nierwetberg, and A. Zacherl, *Accelerated diffusion in josephson junctions and related chaotic systems*, Physical Review Letters, 54 (1985), p. 616.

[123] M. Girvan and M. E. J. Newman, *Community structure in social and biological networks*, Proceedings of the National Academy of Sciences, 99 (2002), pp. 7821–7826.

[124] D. Gleich, *Pagerank beyond the web*, SIAM Review, 57 (2015), pp. 321–363.

[125] P. M. Gleiser and L. Danon, *Community structure in jazz*, Advances in Complex Systems, 6 (2003), pp. 565–573.

[126] B. Gnedenko and A. Kolmogorov, *Independent Random Variables*, Cambridge, Massachusetts: Addison-Wesley, 1954.

[127] J. I. Gold and M. N. Shadlen, *The neural basis of decision making*, Annual Review of Neuroscience, 30 (2007).

[128] N. Goldenfeld and L. P. Kadanoff, *Simple lessons from complexity*, Science, 284 (1999), pp. 87–89.

[129] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, *Understanding individual human mobility patterns*, Nature, 453 (2008), pp. 779–782.

[130] P. Grassberger, *On the critical behavior of the general epidemic process and dynamical percolation*, Mathematical Biosciences, 63 (1983), pp. 157–172.

[131] E. J. Gumbel, *Les valeurs extrêmes des distributions statistiques*, Annales de l'Institut Henri Poincaré, 5 (1935), pp. 115–158.

[132] O. Hallatschek and D. S. Fisher, *Acceleration of evolutionary spread by long-range dispersal*, Proceedings of the National Academy of Sciences, 111 (2014), pp. E4911–E4919.

[133] M. E. Halloran, K. Auranen, S. Baird, N. E. Basta, S. E. Bellan, R. Brookmeyer, B. S. Cooper, V. DeGruttola, J. P. Hughes, J. Lessler, E. T. Lofgren, I. M. Longini, J.-P. Onnela, B. Özler, G. R. Seage, T. A. Smith, A. Vespignani, E. Vynnycky, and M. Lipsitch, *Simulations for designing and interpreting intervention trials in infectious diseases*, BMC Medicine, 15 (2017), p. 223.

[134] T. E. Harris, *Contact interactions on a lattice*, The Annals of Probability, (1974), pp. 969–988.

[135] K.-i. Hashimoto, *Zeta functions of finite graphs and representations of p-adic groups*, Automorphic Forms and Geometry of Arithmetic Varieties, (1989), pp. 211–280.

[136] W. K. Hastings, *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*, Oxford University Press, 1970.

[137] G. Hayrapetyan, F. Iannelli, J. Lekscha, V. Morozov, R. Netz, and Y. S. Mamasakhlisov, *Reentrant melting of rna with quenched sequence randomness*, Physical Review Letters, 113 (2014), p. 068101.

[138] M. Henkel, H. Hinrichsen, and S. Lübeck, *Non-Equilibrium Phase Transitions: Volume 1: Absorbing Phase Transitions*, Theoretical and Mathematical Physics, Springer Netherlands, 2009.

[139] H. W. Hethcote, *The mathematics of infectious diseases*, SIAM Review, 42 (2000), pp. 599–653.

[140] C. A. Hidalgo, *Why Information Grows: The Evolution of Order, From Atoms To Economies*, Penguin Books, 2016.

[141] C. A. Hidalgo and R. Hausmann, *The building blocks of economic complexity*, Proceedings of the National Academy of Sciences, 106 (2009), pp. 10570–10575.

[142] P. Holme, *Temporal network structures controlling disease spreading*, Physical Review E, 94 (2016), p. 022305.

[143] P. Holme and J. Saramäki, *Temporal networks*, Physics Reports, 519 (2012), pp. 97–125.

[144] K. Huang, *Statistical Mechanics, 2nd Edition*, John Wiley & Sons, 1987.

[145] L. Hufnagel, D. Brockmann, and T. Geisel, *Forecast and control of epidemics in a globalized world*, Proceedings of the National Academy of Sciences of the United States of America, 101 (2004), pp. 15124–15129.

[146] B. D. Hughes, *Random Walks and Random Environments*, Oxford University Press, 1995.

[147] V. Hugo, *Les Misérables*, Simon and Schuster, 2013.

[148] F. Iannelli, A. Koher, D. Brockmann, P. Hövel, and I. M. Sokolov, *Effective distances for epidemics spreading on complex networks*, Physical Review E, 95 (2017), p. 012313.

[149] F. Iannelli, M. S. Mariani, and I. M. Sokolov, *Influencers identification in complex networks through reaction-diffusion dynamics*, Phys. Rev. E, 98 (2018), p. 062302.

[150] F. Iannelli, I. M. Sokolov, and F. Thiel, *Reaction-diffusion on random spatial networks with scale-free jumping rates via effective medium theory*, Phys. Rev. E, 98 (2018), p. 032313.

[151] S. Ioos, H.-P. Mallet, I. L. Goffart, V. Gauthier, T. Cardoso, and M. Herida, *Current zika virus epidemiology and recent epidemics*, Medecine et Maladies Infectieuses, 44 (2014), pp. 302 – 307.

[152] L. P. Kadanoff, *Scaling laws for ising models near tc*, Physics Physique Fizika, 2 (1966), p. 263.

[153] L. Katz, *A new status index derived from sociometric analysis*, Psychometrika, 18 (1953), pp. 39–43.

[154] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, vol. 356, van Nostrand Princeton, NJ, 1960.

[155] D. Kempe, J. Kleinberg, and É. Tardos, *Maximizing the spread of influence through a social network*, in Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2003, pp. 137–146.

[156] J. O. Kephart, G. B. Sorkin, D. M. Chess, and S. R. White, *Fighting computer viruses*, Scientific American, 277 (1997), pp. 88–93.

[157] J. O. Kephart, S. R. White, and D. M. Chess, *Computers and epidemiology*, IEEE Spectrum, 30 (1993), pp. 20–26.

[158] S. Kirkpatrick, *Percolation and conduction*, Reviews of Modern Physics, 45 (1973), pp. 574–588.

[159] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, *Identification of influential spreaders in complex networks*, Nature Physics, 6 (2010), pp. 888–893.

[160] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, *Multilayer networks*, Journal of Complex Networks, 2 (2014), pp. 203–271.

[161] I. Kivimaki, M. Shimbo, and M. Saerens, *Developments in the theory of randomized shortest paths with a comparison of graph node distances*, Physica A, 393 (2014), pp. 600 – 616.

[162] J. Klafter, M. F. Shlesinger, and G. Zumofen, *Beyond brownian motion*, Physics Today, 49 (1996), pp. 33–39.

[163] J. Klafter and I. M. Sokolov, *First Steps in Random Walks: From Tools to Applications*, Oxford University Press, 2011.

[164] R. Klages, G. Radons, and I. M. Sokolov, *Anomalous Transport: Foundations and Applications*, John Wiley & Sons, 2008.

[165] D. J. Klein and M. Randić, *Resistance distance*, Journal of Mathematical Chemistry, 12 (1993), pp. 81–95.

[166] D. E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing*, vol. 37, Addison-Wesley Reading, 1993.

[167] V. E. Krebs, *Mapping networks of terrorist cells*, Connections, 24 (2002), pp. 43–52.

[168] D. Krioukov, M. Kitsak, R. S. Sinkovits, D. Rideout, D. Meyer, and M. Boguñá, *Network cosmology*, Scientific Reports, 2 (2012), p. 793.

[169] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, *Spectral redemption in clustering sparse networks*, Proceedings of the National Academy of Sciences, 110 (2013), pp. 20935–20940.

[170] R. Kubo, M. Toda, and N. Hashitsume, *Statistical Physics II: Nonequilibrium Statistical Mechanics*, Springer Science & Business Media, 2012.

[171] R. Lambiotte and M. Rosvall, *Ranking and clustering of nodes in networks with smart teleportation*, Physical Review E, 85 (2012), p. 056107.

[172] G. F. Lawler and V. Limic, *Random Walk: A Modern Introduction*, Cambridge University Press, 2010.

[173] G. Lawyer, *Understanding the influence of all nodes in a network*, Scientific Reports, 5 (2015), p. 8665.

[174] H. H. Lentz, T. Selhorst, and I. M. Sokolov, *Unfolding accessibility provides a macroscopic approach to temporal networks*, Physical Review Letters, 110 (2013), p. 118701.

[175] H. H. K. Lentz, T. Selhorst, and I. M. Sokolov, *Spread of infectious diseases in directed and modular metapopulation networks*, Physical Review E, 85 (2012), p. 066111.

[176] J. Leskovec, L. A. Adamic, and B. A. Huberman, *The dynamics of viral marketing*, ACM Transactions on the Web (TWEB), 1 (2007), p. 5.

[177] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, *Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters*, Internet Mathematics, 6 (2009), pp. 29–123.

[178] J. Leskovec and J. J. Mcauley, *Learning to discover social circles in ego networks*, in Advances in Neural Information Processing Systems, 2012, pp. 539–547.

[179] H. Liao, M. S. Mariani, M. Medo, Y.-C. Zhang, and M.-Y. Zhou, *Ranking in evolving complex networks*, Physics Reports, 689 (2017), pp. 1–54.

[180] J.-G. Liu, Z.-M. Ren, and Q. Guo, *Ranking the spreading influence in complex networks*, Physica A: Statistical Mechanics and its Applications, 392 (2013), pp. 4154–4159.

[181] R. Livi, G. Parisi, S. Ruffo, and A. Vulpiani, *Il computer: da abaco veloce a strumento concettuale*, Il Ponte, (1986), pp. 41–55.

[182] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou, *Vital nodes identification in complex networks*, Physics Reports, 650 (2016), pp. 1–63.

[183] L. Lü, T. Zhou, Q.-M. Zhang, and H. E. Stanley, *The h-index of a network node and its relation to degree and coreness*, Nature Communications, 7 (2016), p. 10168.

[184] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, *The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations*, Behavioral Ecology and Sociobiology, 54 (2003), pp. 396–405.

[185] D. P. Maki and M. Thompson, *Mathematical Models and Applications: With Emphasis On the Social Life, and Management Sciences*, Englewood Cliffs, N.J., Prentice-Hall, 1973.

[186] R. Mancinelli, D. Vergni, and A. Vulpiani, *Front propagation in reactive systems with anomalous diffusion*, Physica D: Nonlinear Phenomena, 185 (2003), pp. 175–195.

[187] B. B. Mandelbrot, *The Fractal Geometry of Nature*, W. H. Freeman, 1983.

[188] S. Manna, *Two-state model of self-organized criticality*, Journal of Physics A: Mathematical and General, 24 (1991), p. L363.

[189] R. N. Mantegna and H. E. Stanley, *Introduction to Econophysics: Correlations and Complexity in Finance*, Cambridge University Press, 2000.

[190] A. Mårell, J. P. Ball, and A. Hofgaard, *Foraging and movement paths of female reindeer: insights from fractal analysis, correlated random walks, and lévy flights*, Canadian Journal of Zoology, 80 (2002), pp. 854–865.

[191] J. Marro and R. Dickman, *Nonequilibrium Phase Transitions in Lattice Models*, Cambridge University Press, 2005.

[192] T. Martin, X. Zhang, and M. E. J. Newman, *Localization and centrality in networks*, Physical Review E, 90 (2014), p. 052808.

[193] L. Meyers, *Contact network epidemiology: Bond percolation applied to infectious disease prediction and control*, Bulletin of the American Mathematical Society, 44 (2007), pp. 63–86.

[194] M. Mézard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*, vol. 9, World Scientific Publishing Company, 1987.

[195] R. Michalski, S. Palus, and P. Kazienko, *Matching Organizational Structure and Social Network Extracted from Email Communication*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 197–206.

[196] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, *Measurement and Analysis of Online Social Networks*, in Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07), San Diego, CA, October 2007.

[197] J. D. Murray, *Mathematical Biology II – Spatial Models and Biomedical Applications*, Springer-Verlag New York Incorporated New York, 2001.

[198] M. E. J. Newman, *Spread of epidemic disease on networks*, Physical Review E, 66 (2002), p. 016128.

[199] M. E. J. Newman, *Power laws, pareto distributions and zipf's law*, Contemporary Physics, 46 (2005), pp. 323–351.

[200] M. E. J. Newman, *Finding community structure in networks using the eigenvectors of matrices*, Physical Review E, 74 (2006).

[201] M. E. J. Newman, *Networks: An Introduction*, Oxford University Press, 2010.

[202] M. E. J. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics*, Oxford University Press, 1999.

[203] G. Nicolis and C. Nicolis, *Foundations of Complex Systems: Nonlinear Dynamics, Statistical Physics, Information and Prediction*, Wspc, 2007.

[204] G. Nicolis and I. Prigogine, *Self-Organization in Nonequilibrium Systems*, John Wiley & Sons, New York, 1977.

[205] A. Nordsieck, W. Lamb Jr, and G. Uhlenbeck, *On the theory of cosmic-ray showers i the furry model and the fluctuation problem*, Physica, 7 (1940), pp. 344–360.

[206] J. R. Norris, *Markov Chains*, Cambridge University Press, 1998.

[207] J. Novak, *Polya's random walk theorem*, The American Mathematical Monthly, 121 (2014), pp. 711–716.

[208] Z. N. Oltvai and A.-L. Barabási, *Life's complexity pyramid*, Science, 298 (2002), pp. 763–764.

[209] L. Onsager, *Crystal statistics i. a two-dimensional model with an order-disorder transition*, Physical Review, 65 (1944), p. 117.

[210] L. Page, S. Brin, R. Motwani, and T. Winograd, *The pagerank citation ranking: Bringing order to the web.*, tech. rep., Stanford InfoLab, 1999.

[211] A. Pagnani, G. Parisi, and F. Ricci-Tersenghi, *Glassy transition in a disordered model for the rna secondary structure*, Physical Review Letters, 84 (2000), p. 2026.

[212] G. Parisi, *Statistical Field Theory*, Addison-Wesley Pub. Co., 1988.

[213] R. Pastor-Satorras and C. Castellano, *Topological structure and the h index in complex networks*, Physical Review E, 95 (2017), p. 022301.

[214] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, *Epidemic processes in complex networks*, Reviews of Modern Physics, 87 (2015), pp. 925–979.

[215] R. Pastor-Satorras and A. Vespignani, *Epidemic spreading in scale-free networks*, Physical Review Letters, 86 (2001), pp. 3200–3203.

[216] R. Pastor-Satorras and A. Vespignani, *Epidemic dynamics in finite size scale-free networks*, Physical Review E, 65 (2002), p. 035108.

[217] R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach*, Cambridge University Press, 2004.

[218] K. Pearson, *The problem of the random walk*, Nature, 72 (1905), p. 342.

[219] S. Pei, L. Muchnik, J. S. Andrade Jr, Z. Zheng, and H. A. Makse, *Searching for superspreaders of information in real-world social media*, Scientific Reports, 4 (2014), p. 5547.

[220] T. P. Peixoto, *Efficient monte carlo and greedy heuristic for the inference of stochastic block models*, Physical Review E, 89 (2014), p. 012804.

[221] T. P. Peixoto, *Hierarchical block structures and high-resolution model selection in large networks*, Physical Review X, 4 (2014), p. 011047.

[222] A. Pentland, *Social Physics: How Good Ideas Spread-The Lessons from a New Science*, Penguin, 2014.

[223] M. E. Peskin and D. V. Schroeder, *Quantum field theory*, The Advanced Book Program, Perseus Books Reading, Massachusetts, (1995).

[224] G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. J. Hellyer, and F. Vaccarino, *Homological scaffolds of brain functional networks*, Journal of The Royal Society Interface, 11 (2014).

[225] A. P. Y. Piontti, M. F. D. C. Gomes, N. Samay, N. Perra, and A. Vespignani, *The infection tree of global epidemics*, Network Science, 2 (2014), pp. 132–137.

[226] C. Poletto, M. F. Gomes, A. P. y Piontti, L. Rossi, L. Bioglio, D. L. Chao, I. M. Longini, M. E. Halloran, V. Colizza, and A. Vespignani, *Assessing the impact of travel restrictions on international spread of the 2014 west african ebola epidemic*, Eurosurveillance, 19 (2014).

[227] G. Pólya, *Über eine aufgabe der wahrscheinlichkeitsrechnung betreffend die irrfahrt im straßennetz*, Mathematische Annalen, 84 (1921), pp. 149–160.

[228] F. Radicchi and C. Castellano, *Leveraging percolation theory to single out influential spreaders in networks*, Physical Review E, 93 (2016), p. 062314.

[229] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, *Defining and identifying communities in networks*, Proceedings of the National Academy of Sciences, 101 (2004), pp. 2658–2663.

[230] G. Ramos-Fernández, J. L. Mateos, O. Miramontes, G. Cocho, H. Larralde, and B. Ayala-Orozco, *Lévy walk patterns in the foraging movements of spider monkeys (ateles geoffroyi)*, Behavioral Ecology and Sociobiology, 55 (2004), pp. 223–230.

[231] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer, *Truthy: mapping the spread of astroturf in microblog streams*, in Proceedings of the 20th International Conference Companion on World Wide Web, ACM, 2011, pp. 249–252.

[232] S. Redner, *A Guide to First-Passage Processes*, Cambridge University Press, 2001.

[233] S. Riley, *Large-scale spatial-transmission models of infectious disease*, Science, 316 (2007), pp. 1298–1301.

[234] L. E. Rocha, F. Liljeros, and P. Holme, *Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts*, PLOS Computational Biology, 7 (2011), p. e1001109.

[235] M. Rosvall, A. V. Esquivel, A. Lancichinetti, J. D. West, and R. Lambiotte, *Memory in network flows and its effects on spreading dynamics and community detection*, Nature Communications, 5 (2014).

[236] L. A. Rvachev and I. M. Longini, *A mathematical model for the global spread of influenza*, Mathematical Biosciences, 75 (1985), pp. 3–22.

[237] G. Sabidussi, *The centrality index of a graph*, Psychometrika, 31 (1966), pp. 581–603.

[238] M. Salathé, *Digital epidemiology: what is it, and where is it going?*, Life Sciences, Society and Policy, 14 (2018), p. 1.

[239] M. Salathé, L. Bengtsson, T. J. Bodnar, D. D. Brewer, J. S. Brownstein, C. Buckee, E. M. Campbell, C. Cattuto, S. Khandelwal, P. L. Mabry, and A. Vespignani, *Digital epidemiology*, PLOS Computational Biology, 8 (2012), p. e1002616.

[240] M. Salathé and S. Khandelwal, *Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control*, PLOS Computational Biology, 7 (2011), pp. 1–7.

[241] L. M. Sander, C. P. Warren, I. M. Sokolov, C. Simon, and J. Koopman, *Percolation on heterogeneous networks as a model for epidemics*, Mathematical Biosciences, 180 (2002), pp. 293–305.

[242] K.-I. Sato, *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press, Cambridge, 1999.

[243] I. Scholtes, N. Wider, R. Pfitzner, A. Garas, C. J. Tessone, and F. Schweitzer, *Causality-driven slow-down and speed-up of diffusion in non-markovian temporal networks*, Nature Communications, 5 (2014), p. 5024.

[244] M. Schroeder, *Fractals, Chaos, Power Laws: Minutes From an Infinite Paradise*, Courier Corporation, 2009.

[245] F. Schweitzer, *Sociophysics*, Physics Today, 71 (2018), p. 40.

[246] S. B. Seidman, *Internal cohesion of ls sets in graphs*, Social Networks, 5 (1983), pp. 97–107.

[247] S. B. Seidman, *Network structure and minimum degree*, Social Networks, 5 (1983), pp. 269–287.

[248] M. A. Serrano, D. Krioukov, and M. Boguná, *Self-similarity of complex networks and hidden metric spaces*, Physical Review Letters, 100 (2008), p. 078701.

[249] D. Sherrington and S. Kirkpatrick, *Solvable model of a spin-glass*, Physical Review Letters, 35 (1975), pp. 1792–1796.

[250] C. Song, S. Havlin, and H. A. Makse, *Self-similarity of complex networks*, Nature, 433 (2005), p. 392.

[251] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*, CRC Press, 2018.

[252] P. Suárez-Serrato, M. E. Roberts, C. Davis, and F. Menczer, *On the influence of social bots in online protests*, in International Conference on Social Informatics, Springer, 2016, pp. 269–278.

[253] K. Symanzik, *Small distance behaviour in field theory and power counting*, Communications in Mathematical Physics, 18 (1970), pp. 227–246.

[254] A. Tacchella, M. Cristelli, G. Caldarelli, A. Gabrielli, and L. Pietronero, *A new metrics for countries' fitness and products' complexity*, Scientific Reports, 2 (2012), p. 723.

[255] V. Tejedor, O. Bénichou, and R. Voituriez, *Global mean first-passage times of random walks on complex networks*, Physical Review E, 80 (2009), p. 065104.

[256] F. Thiel and I. M. Sokolov, *Effective-medium approximation for lattice random walks with long-range jumps*, Physical Review E, 94 (2016), p. 012135.

[257] V. M. Tikhomirov, *A study of the diffusion equation with increase in the amount of substance, and its application to a biological problem*, in Selected Works of AN Kolmogorov, Springer, 1991, pp. 242–270.

[258] M. Tizzoni, P. Bajardi, C. Poletto, J. J. Ramasco, D. Balcan, B. Gonçalves, N. Perra, V. Colizza, and A. Vespignani, *Real-time numerical forecast of global epidemic spreading: case study of 2009 a/h1n1pdm*, BMC Medicine, 10 (2012), p. 165.

[259] M. Tribus, *Thermostatics and Thermodynamics*, Center for Advanced Engineering Study, Massachusetts Institute of Technology, 1970.

[260] L. G. Valiant, *The complexity of enumeration and reliability problems*, SIAM Journal on Computing, 8 (1979), pp. 410–421.

[261] W. Van den Broeck, C. Gioannini, B. Gonçalves, M. Quaggiotto, V. Colizza, and A. Vespignani, *The gleamviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale*, BMC Infectious Diseases, 11 (2011), p. 37.

[262] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry*, Elsevier, 1992.

[263] P. Van Mieghem, *Graph Spectra for Complex Networks*, Cambridge University Press, 2010.

[264] A. Vespignani, *Predicting the behavior of techno-social systems*, Science, 325 (2009), pp. 425–428.

[265] A. Vespignani, *Modelling dynamical processes in complex socio-technical systems*, Nature Physics, 8 (2012), p. 32.

[266] C. Viboud, O. N. Bjørnstad, D. L. Smith, L. Simonsen, M. A. Miller, and B. T. Grenfell, *Synchrony, waves, and spatial hierarchies in the spread of influenza*, Science, 312 (2006), pp. 447–451.

[267] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet, *Novel type of phase transition in a system of self-driven particles*, Physical Review Letters, 75 (1995), p. 1226.

[268] G. M. Viswanathan, V. Afanasyev, S. V. Buldyrev, E. J. Murphy, P. A. Prince, and H. E. Stanley, *Lévy flight search patterns of wandering albatrosses*, Nature, 381 (1996), pp. 413–415.

[269] G. M. Viswanathan, S. V. Buldyrev, S. Havlin, M. Da Luz, E. Raposo, and H. E. Stanley, *Optimizing the success of random searches*, Nature, 401 (1999), p. 911.

[270] C.-J. Wang, L. Wu, J. Zhang, and M. Janssen, *The hidden geometry of attention diffusion*, arXiv preprint arXiv:1501.06552, (2015).

[271] C. P. Warren, L. M. Sander, and I. M. Sokolov, *Firewalls, disorder, and percolation in epidemics*, arXiv preprint cond-mat/0106450, (2001).

[272] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, vol. 8, Cambridge University Press, 1994.

[273] D. J. Watts, *A simple model of global cascades on random networks*, Proceedings of the National Academy of Sciences, 99 (2002), pp. 5766–5771.

[274] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature, 393 (1998), pp. 440–442.

[275] R. J. Williams and N. D. Martinez, *Limits to trophic levels and omnivory in complex food webs: theory and data*, The American Naturalist, 163 (2004), pp. 458–468.

[276] K. G. Wilson, *Renormalization group and critical phenomena i. renormalization group and the kadanoff scaling picture*, Physical Review B, 4 (1971), p. 3174.

[277] T. A. Witten and L. M. Sander, *Diffusion-limited aggregation, a kinetic critical phenomenon*, Physical Review Letters, 47 (1981), p. 1400.

[278] J. Yang and J. Leskovec, *Defining and evaluating network communities based on ground-truth*, Knowledge and Information Systems, 42 (2015), pp. 181–213.

[279] Y. Yang, J. D. Sugimoto, M. E. Halloran, N. E. Basta, D. L. Chao, L. Matrajt, G. Potter, E. Kenah, and I. M. Longini, *The transmissibility and control of pandemic influenza a (h1n1) virus*, Science, 326 (2009), pp. 729–733.

[280] W. W. Zachary, *An information flow model for conflict and fission in small groups*, Journal of Anthropological Research, 33 (1977), pp. 452–473.

[281] A. Zeng and C.-J. Zhang, *Ranking spreaders by decomposing complex networks*, Physics Letters A, 377 (2013), pp. 1031–1035.

[282] B. Zhang, R. Liu, D. Massey, and L. Zhang, *Collecting the Internet AS-level topology*, SIGCOMM Computer Communication Review, 35 (2005), pp. 53–61.

[283] F. Zhang, *Matrix Theory: Basic Results and Techniques*, Springer Science & Business Media, 2011.

[284] Z.-K. Zhang, C. Liu, X.-X. Zhan, X. Lu, C.-X. Zhang, and Y.-C. Zhang, *Dynamics of information diffusion and its applications on complex networks*, Physics Reports, 651 (2016), pp. 1–34.

[285] J. Zinn-Justin, *Quantum Field Theory and Critical Phenomena*, Oxford University Press, 2002.

[286] R. Zwanzig, *Nonequilibrium Statistical Mechanics*, Oxford University Press, 2001.

Selbständigkeitserklärung

Ich erkläre, dass ich die Dissertation selbständig und nur unter Verwendung der von mir gemäß § 7 Abs. 3 der Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät, veröffentlicht im Amtlichen Mitteilungsblatt der Humboldt-Universit at zu Berlin Nr. 126/2014 am 18.11.2014 angegebenen Hilfsmittel angefertigt habe.

Berlin, den 25. Februar 2019 Flavio Iannelli